КЛАССИЧЕСКОЕ УНИВЕРСИТЕТСКОЕ ИЗДАНИЕ

Серия основана в 2010 году



КЛАССИЧЕСКОЕ УНИВЕРСИТЕТСКОЕ ИЗДАНИЕ

О. В. Терещенко Н. В. Курилович
Е. И. Князева

МНОГОМЕРНЫЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ В СОЦИАЛЬНЫХ НАУКАХ

*Допущено
Министерством образования Республики Беларусь
в качестве учебного пособия для студентов
учреждений высшего образования
по социально-гуманитарным специальностям*



МИНСК
БГУ
2012

УДК 303.71(075.8)
ББК 60.в631.8я73
Т35

Р е ц е н з е н т ы:
кафедра экономической социологии БГЭУ
(заведующий кафедрой кандидат философских наук,
доцент *В. Я. Кочергин*);
доктор социологических наук, профессор *С. А. Шавель*

Терещенко, О. В.

Т35 Многомерный статистический анализ данных в социальных науках : учеб. пособие / О. В. Терещенко, Н. В. Курилович, Е. И. Князева. – Минск : БГУ, 2012. – 239 с. : ил. – (Классическое университетское издание).

ISBN 978-985-518-711-1.

Рассмотрены многомерные методы статистического анализа данных. Основное внимание уделено моделям анализа данных, условиям их применения, а также особенностям представления данных и интерпретации результатов.

Предназначено для студентов учреждений высшего образования, обучающихся по социально-гуманитарным специальностям.

УДК 303.71(075.8)
ББК 60.в631.8я73

ISBN 978-985-518-711-1

Уважаемые читатели!

Серия «Классическое университетское издание» была основана в 2010 году к 90-летию Белорусского государственного университета. Путь, который прошло наше учебное заведение в своем развитии, свидетельствует о становлении в нем собственной академической и научной традиции. Несомненно, опыт и знания, аккумулированные в стенах БГУ, являются не только предметом нашей гордости, но и достоянием всего белорусского общества. Одна из целей предлагаемой серии – сделать это достояние как можно более открытым и доступным.

Белорусский государственный университет всегда славился академичностью и фундаментальностью в подготовке специалистов. Однако сегодня этого уже недостаточно. От выпускника требуется умение быстро включаться в непосредственную практическую работу, которой свойствен синтез нескольких форм деятельности: собственно производственной, исследовательской, проектно-разработческой. В выигрыше в конечном итоге окажется тот, кто сегодня научится более эффективно создавать и применять знания, оперативно изменять технологии, совершенствовать и радикально трансформировать накопленный опыт. Вот почему совмещение преимуществ фундаментального и прагматического образования стало основой инновационно ориентированной подготовки будущих специалистов в нашем университете.

Серия отражает многолетний опыт научно-педагогической, методической и издательской работы БГУ. Ее цель – предста-

вить модель учебного текста, которая в своей структуре содержит набор программ образовательно-научно-производственной деятельности будущих специалистов. Реализация этой модели позволит обеспечить универсализм выпускника, его способность к эффективному решению важных задач, стоящих перед Республикой Беларусь на национальном и международном уровне.

Классическое университетское издание, являя собой сплав научной и педагогической мысли, призвано формировать особую культуру знания – передового и доступного, теоретического и практического, общекультурного и специализированного. Словом, такого знания, которое будет работать.

Книги этой серии должны стать образцом научно-методического обеспечения современного образовательного процесса в высшей школе, утвердить ведущую роль нашего университета в качестве национального научно-методического центра Республики Беларусь.

Надеемся, что серия «Классическое университетское издание» состоится и как одно из слагаемых особой культурно-образовательной среды БГУ, которая будет способствовать интеллектуальному росту и творческой созидательной деятельности наших студентов.

Ректор Белорусского государственного университета академик НАН Беларуси, профессор *С. В. Абламейко*

ВВЕДЕНИЕ

Цель компьютерной обработки –
понимание, а не число.
Р. Хемминг

Многомерные методы статистического анализа данных появились еще в начале XX в. Однако из-за большого объема и сложности вычислений они получили широкое распространение только благодаря созданию компьютеров, особенно персональных, с «дружественными» операционными системами и пользовательскими интерфейсами.

Концепции данного учебного пособия способствовала дискуссия 1970–80-х гг. о том, больше пользы или вреда приносит развитие интерфейсов статистического программного обеспечения, с одной стороны, расширяющих возможности доступа исследователям, не имеющим специальной математической подготовки, к сложным статистическим методам, с другой – грозящих профанацией их применения некомпетентными пользователями и получением результатов, достоверность, надежность и репрезентативность которых плохо поддаются контролю. Прошедшие десятилетия показали, что коммерциализацию программных средств и связанную с ней демократизацию доступа к статистическим методам не остановить и что сбылись связанные с этим не только надежды, но и опасения.

В учебном пособии основное внимание уделяется не вычислениям (что с распространением персональных компьютеров утратило актуальность) и не использованию компьютерного программного обеспечения (в этой области в последние годы появилось достаточно много литературы[1]), а моделям анализа данных, условиям их применения, особенностям представления данных и интерпретации результатов. В большинстве случаев мы опускаем строгие математические решения, лежащие в основе моделей, популярных в социально-гуманитарных науках, но приводим необходимые ссылки, позволяющие студенту получить более точную и подробную информацию.

[1] *Бессокирная Г. П.* Анализ социологических данных с помощью SPSS : Обзор учебной литературы // Социология : 4М. 2008. № 26. С. 68–175.

Многообразие методов многомерного статистического анализа данных определяется как решаемыми задачами, так и особенностями исходных данных, которые в социологии и смежных науках отличаются разнообразием.

Во-первых, в исходных данных *преобладают категориальные переменные*, для которых методы классической статистики не предназначены; поэтому мы обсуждаем некоторые специальные модели и техники, ориентированные на применение таких переменных (логистическая регрессия, *dummy*-кодирование и др.).

Во-вторых, данные бывают атрибутивными и реляционными. *Атрибутивные данные* характеризуют объекты из изучаемой статистической совокупности – генеральной или выборочной. Их структурируют по двум основным критериям – «объект» и «переменная» – и представляют в виде соответствующей матрицы данных. Примерами атрибутивных переменных могут служить пол, возраст, статус, удовлетворенность работой и т. п. В панельных исследованиях, когда переменные измеряются на одних и тех же объектах несколько раз, может быть введен третий критерий – «время измерения». *Реляционные данные* характеризуют структуру отношений и / или связей между объектами, составляющими выборку исследования. Методы анализа реляционных данных восходят к социометрии и анализу социальных сетей. Социометрическая матрица является классическим примером реляционных данных, однако это не единственный их вид. Кроме социометрии и анализа социальных сетей, подобные данные могут иметь место, например, при использовании методов парных и множественных сравнений в многомерном шкалировании.

В-третьих, вид данных далеко не всегда сводится к классической матрице «объект – переменная». Помимо принятого в экономике и официальной статистике временного измерения, превращающего матрицу данных в «куб», бывают матрицы одинакового вида, полученные из нескольких источников (например, в исследованиях методом семантического дифференциала).

За каждым математическим методом стоит определенная модель изучаемого с его помощью явления. Так, используя регрессионный анализ, мы полагаем, что значения зависимой переменной линейно связаны со значениями независимой переменной. Используя дисперсионный анализ, делаем соответствующие предположения о факторах, формирующих значение зависимой переменной для каждой из рассматриваемых «клеток» плана эксперимента – совокупностей объектов, обладающих определенным набором значений факторов. Применяя методы классификации, предполагаем, что искомым классам соответствуют определенные совокупности классифицируемых объектов, которые могут рассматриваться как точки некоторого признакового пространства и т. п. Выбор модели такого рода, по существу, и означает выбор математического метода.

В основе большинства статистических методов лежит идея «сжатия» данных до одного или нескольких чисел, например, при вычислении среднего арифметического или построении частотных распределений. В случае многомерной статистики речь чаще всего идет об агрегировании столбцов и / или

строк исходной матрицы данных[1]. Представление данных в виде матрицы «объект – переменная» позволяет решать два вида задач: анализ взаимосвязей между переменными – столбцами матрицы (исследование структуры связей, снижение размерности) и выявление сходства между объектами – строками матрицы (классификация). В обоих случаях целью является «сжатие» информации. Методы снижения размерности«сжимают» матрицу по столбцам, выделяя группы связанных друг с другом переменных, за каждой из которых усматривается действие одного латентного фактора. Методы классификации, объединяя в кластеры схожие между собой объекты, «сжимают» матрицу данных по строкам.

Что касается методов исследования причинных связей, они также базируются на изучении структуры связей между зависимыми и независимыми переменными, однако «сжатие» информации здесь производится иным способом – через построение математической модели.

Различия в постановке задач являются основным критерием классификации методов анализа данных, в соответствии с которым принято выделять три группы методов многомерной статистики.

Методы исследования причинных связей между переменными. При решении данной задачи между переменными устанавливаются отношения причинности: определяются независимые переменные, измеряющие причины, и зависимые переменные, измеряющие следствия. В анализе причинных связей можно выделить три подзадачи: анализ структуры причинных связей, их статистическое моделирование, прогнозирование и объяснение изменений зависимых переменных при помощи построенных моделей. Для решения данной группы задач используются регрессионные методы (множественная линейная регрессия, логистическая регрессия и др.), а также логически близкие к ним путевой анализ и моделирование линейно-структурными уравнениями.

Методы исследования структуры корреляционных связей между переменными. Указанная задача может решаться как сама по себе (определение подмножеств переменных, тесно связанных между собой), так и с целью сокращения количества используемых переменных (снижения размерности) без существенной потери информации. Для этого используются методы главных компонент, факторного анализа, кластерного анализа переменных, многомерного шкалирования.

Методы исследования структуры изучаемой совокупности объектов (*генеральной или выборочной*). В рамках этой задачи осуществляется классификация изучаемых объектов на основе одного из двух подходов. Первый подход состоит в группировке объектов по нескольким критериям одновременно. Второй основан на вычислении расстояний между объектами в качестве показателя степени различий между ними: чем больше различий, например, между ответами респондентов на вопросы анкеты, тем больше расстояния и меньше шансов для объектов попасть в один класс. На вычислении расстояний базируются методы кластерного анализа объектов, дискриминантного анализа, а

---

[1] *Толстова Ю.Н.* Измерение в социологии : курс лекций. М., 1998. С. 137.

также некоторые модели такого метода снижения размерности, как многомерное шкалирование.

Особое внимание в учебном пособии уделено двум методам, не получившим пока широкого распространения, но, на наш взгляд, весьма перспективным. Эти методы выделены по критерию редко используемых видов данных. *Анализ социальных сетей* предназначен для работы с реляционными данными. В качестве единицы анализа в нем выступает не объект из выборки, и даже не актор социальной сети, а разнообразные связи и отношения между двумя акторами. *Когортный анализ* применяется в исследованиях социальной динамики для анализа данных повторных исследований, как мониторинговых, так и панельных. Это единственный метод, позволяющий корректно разделить в социодинамических процессах «эффект возраста», связанный с возрастными изменениями личностных качеств, жизненным опытом, доступностью ресурсов, «эффект когорты», соотносящийся с условиями ее возникновения, становления, социализации, и «эффект времени», обусловленный переменами, происходящими в макросреде, – экономическими, политическими, культурными, информационными и др.

Несколько слов необходимо сказать о приложении к учебному пособию. В него вошли практически недоступные современному читателю статьи ведущих российских и белорусских аналитиков, вышедшие в первой половине 1990-х гг. весьма ограниченными тиражами. В этих статьях глубокий анализ данных сочетается с подробным обсуждением применяемых методов. Позднее подобные работы не появлялись в русскоязычной социологической литературе, т. к. начиная со второй половины 1990-х гг. «тонкие» особенности использования аналитических методов стали считаться ноу-хау исследовательских (в первую очередь, маркетинговых и политических) организаций. Обсуждение методов анализа данных продолжается исключительно в учебной литературе, где используются, как правило, дидактические примеры, а не примеры, построенные на реальных данных.

Настоящее учебное пособие предназначено для студентов, магистрантов, аспирантов, изучающих методологию количественных исследований и методы статистического анализа данных; преподавателей, научных сотрудников и специалистов, принимающих участие в эмпирических социальных исследованиях, – социологов, психологов, политологов, маркетологов, менеджеров, специалистов по коммуникации. Его изучение предполагает предварительное знакомство с основами прикладной статистики (на уровне одномерных распределений, анализа парных связей, проверки статистических гипотез) и одним из программных средств статистического анализа данных (SPSS, Statistica, R и т. п.).

Издание подготовлено авторским коллективом факультета философии и социальных наук БГУ. Главы 1–3 и 5 написаны О. В. Терещенко, разделы 1.1. и 1.2, а также главы 2 и 3 – Н. В. Курилович, глава 4 – Е. И. Князевой.

Глава 1

КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ

1.1. СТРУКТУРА СВЯЗЕЙ МЕЖДУ ПЕРЕМЕННЫМИ

В зависимости от вкладываемого в исследование смысла связи между переменными могут быть как корреляционными, так и причинными (каузальными). Связь между двумя переменными называется корреляционной, если они рассматриваются как двусторонне взаимодействующие, без выделения причины и следствия. Связь называется причинной, если одна из переменных (зависимая) измеряет следствие, а другая или несколько независимых переменных (предикторов) измеряют одну или несколько причин.

Меры корреляционной связи. Большинство статистических мер связи предназначены для измерения парных корреляционных связей между переменными. Разумеется, они могут использоваться и для причинных связей на тех уровнях анализа, когда их причинно-следственное содержание игнорируется.

Выбор меры связи между двумя переменными зависит, в первую очередь, от уровня их измерения: для двух количественных переменных это коэффициент линейной корреляции Пирсона; для двух порядковых переменных – коэффициенты ранговой корреляции Спирмана и Кендалла; для двух дихотомических переменных – коэффициенты Ф (фи) и Юла; для номинальных переменных с числом градаций более двух – коэффициент Крамера. Если переменные имеют разный уровень измерения, выбирается коэффициент, соответствующий более низкому уровню. Например, если одна переменная является количественной, а вторая порядковой, рекомендуется использовать одну из порядковых мер связи (возможно, количественную переменную придется при этом сгруппи-

ровать в интервалы); если одна из переменных является номинальной, а вторая – порядковой, следует использовать коэффициент Крамера. Формулы для вычисления коэффициентов мы здесь не приводим, т. к., во-первых, они имеются во всех справочных изданиях, начиная с «Рабочей книги социолога»[1] (первое издание вышло в 1977 г.), во-вторых, в многомерной статистике наиболее часто используется коэффициент линейной корреляции Пирсона в силу его универсальности, о чем речь пойдет ниже.

Наибольший интерес в задачах многомерной статистики представляют переменные, связи между которыми обладают «направленностью», т. е. могут трактоваться как «прямые» или «обратные». Понятие направленности может применяться только в двух случаях. Во-первых, когда обе переменные являются количественными и / или порядковыми: связь является прямой, если значения двух переменных одновременно увеличиваются или уменьшаются; обратной – если увеличение значения одной переменной сопровождается уменьшением значения другой. Во-вторых, когда обе переменные являются дихотомическими: связь является прямой, если два фиксируемых свойства объектов чаще встречаются и не встречаются совместно, чем порознь; обратной – если соответствующие свойства чаще встречаются порознь, чем совместно.

Коэффициент корреляции между переменными x_i и x_j обозначается $r_{i,j}$ и обладает следующими свойствами:

1) коэффициент корреляции симметричен ($r_{i,j} = r_{j,i}$);

2) значение коэффициента корреляции находится в пределах $-1 \le r_{i,j} \le 1$ для направленных связей; $0 \le r_{i,j} \le 1$ для ненаправленных связей;

3) $r_{i,j} = 0$, если связи между переменными нет;

4) $r_{i,j} > 0$, если связь является прямой или ненаправленной;

5) $r_{i,j} < 0$, если связь является обратной;

6) $r_{i,j} = \pm 1$, если связь является полной, т. е. по значению одной переменной можно точно определить значение второй.

Идеальным случаем для задач снижения размерности и классификации является использование переменных с одинаковым уровнем измерения: количественных, порядковых (измеренных с использованием шкал Лайкерта[2] и подобных им оценочных шкал с четным или нечетным количеством

[1] Рабочая книга социолога / под общ. ред. Г. В. Осипова. Изд. 5-е. М., 2009. С. 176–190.

[2] Шкала Лайкерта – порядковая оценочная 5-балльная шкала, предполагающая ответы типа «да» – «скорее да» – «ни да, ни нет» – «скорее нет» – «нет». Подробнее см.: *Толстова Ю. Н.* Измерение в социологии : курс лекций. М., 1998. С. 111–112.

чеством градаций, которые могут рассматриваться как квазиинтервальные[1]) или дихотомических. Номинальные переменные, не являющиеся дихотомическими, а также порядковые переменные, которые не могут рассматриваться как квазиинтервальные, используются только в отдельных многомерных статистических моделях с применением специально разработанных для этого техник, которые будут рассмотрены в соответствующих разделах.

В ситуации, когда одновременно должны анализироваться несколько переменных с разным уровнем измерения, рекомендуется использовать коэффициент парной корреляции Пирсона (формула 1.1), универсальный в том смысле, что коэффициент ранговой корреляции Спирмана, коэффициент Ф для двух дихотомических переменных и коэффициент бисериальной корреляции[2], используемый, когда одна переменная количественная, а вторая – дихотомическая, являются полными его аналогами:

$$r_{i,j} = \frac{\sum_{l=1}^{n}\left(x_i(l)-\bar{x}_i\right)\left(x_j(l)-\bar{x}_j\right)}{s_i s_j}, \qquad (1.1)$$

где n – объем выборки; x_i, x_j – переменные с номерами i и j; $x_i(l)$, $x_j(l)$ – значения переменных x_i и x_j для объекта (респондента) из выборки с номером $l\left(l=\overline{1,n}\right)$; $\bar{x}_i$, $\bar{x}_j$ – средние арифметические пременных x_i и x_j; s_i, s_j – стандартные отклонения переменных x_i и x_j.

Матрица корреляций. Для представления структуры связей между переменными используется матрица корреляций. Это квадратная таблица[3], в которой строки и столбцы соответствуют одним и тем же переменным (общее количество переменных будем обозначать буквой k, размерность матрицы $k \times k$). В клетке на пересечении строки с номером i и столбца с номером j указывается значение коэффициента корреляции $r_{i,j}$ для переменных x_i и x_j (табл. 1.1).

[1] Предполагается, что, отвечая на вопрос с использованием шкалы Лайкерта и подобных ей, респондент интуитивно исходит из того, что расстояния между градациями шкалы одинаковы, т. е. разность между ответами «да» и «скорее да» такая же, как между ответами «скорее да» и «ни да, ни нет», и т. п. В этом случае на шкале действует отношение разности (как и на шкале интервалов) и к ней применимы арифметические действия, в частности, вычисление среднего арифметического и дисперсии.

[2] *Шавель С. А.* Бисериальная корреляция // Словарь прикладной социологии / сост. К. В. Шульга ; отв. ред. Г. П. Давидюк. Минск, 1984. С. 11–15.

[3] Таблица является квадратной, если в ней одинаковое количество строк и столбцов.

В матрице корреляций могут использоваться любые меры связи при условии, что все переменные имеют одинаковый уровень измерения. Так, если все переменные количественные, используется коэффициент Пирсона (r), если порядковые – коэффициент Спирмана (r_s) или Кендалла (τ), если дихотомические – коэффициент Ф (фи), если номинальные – коэффициент Крамера (V).

Таблица 1.1

Матрица корреляций ($i, j = 1, k$)

Переменная	x_1	x_2	...	x_j	...	x_k
x_1	1	$r_{1,2}$	...	$r_{1,j}$	...	$r_{1,k}$
x_2	$r_{2,1}$	1	...	$r_{2,j}$	...	$r_{2,k}$
...	...	...	...	...	...	...
x_i	$r_{i,1}$	$r_{i,2}$	...	$r_{i,j}$	...	$r_{i,k}$
...	...	...	...	...	...	...
x_k	$r_{k,1}$	$r_{k,2}$	...	$r_{k,j}$	...	1

Если переменные имеют разный уровень измерения, в матрицу корреляций не должны включаться номинальные переменные, не являющиеся дихотомическими. Для переменных с любыми другими уровнями измерения вычисляется коэффициент корреляции Пирсона с учетом его модификаций для порядковых и дихотомических переменных. При компьютерной обработке данных это происходит автоматически.

Матрица корреляций симметрична относительно главной диагонали ($r_{i,j} = r_{j,i}$), которая полностью состоит из единиц (коэффициент корреляции переменной с самой собой равен 1). Поэтому она может быть представлена в форме верхнего (табл. 1.2) или нижнего треугольника.

Пример 1.1. Влияние внутреннего валового продукта (ВВП) на социально-демографические показатели (европейские страны, 2008 г.)

Переменные:

1. Медианный возраст населения страны.
2. Рождаемость (число родившихся на 1000 жителей).
3. Смертность (число умерших на 1000 жителей).
4. Естественный прирост (разность рождаемости и смертности на 1000 жителей).
5. Детская смертность (до 1 года на 1000 живорожденных).
6. Ожидаемая продолжительность жизни мужчин при рождении.
7. Ожидаемая продолжительность жизни женщин при рождении.
8. Валовой внутренний продукт на душу населения.

Таблица 1.2

Матрица корреляций между социально-демографическими показателями

Переменная	1	2	3	4	5	6	7	8
1. Медианный возраст	1	**-0,74**	0,26	**-0,55**	**-0,78**	0,39	0,53	0,39
2. Рождаемость		1	-0,46	**0,81**	0,50	-0,01	-0,08	0,08
3. Смертность			1	**-0,90**	-0,04	**-0,75**	**-0,63**	**-0,55**
4. Естественный прирост				1	0,50	0,38	0,27	0,40
5. Детская смертность					1	-0,49	**-0,63**	**-0,58**
6. Ожидаемая продолжительность жизни мужчин						1	**0,92**	**0,77**
7. Ожидаемая продолжительность жизни женщин							1	**0,83**
8. ВВП								1

Граф матрицы корреляций. В социологии широко распространены структурные модели. Их используют для анализа структуры связей между переменными, структуры исследуемой статистической совокупности объектов, структуры набора понятий в представлениях респондентов и др. По методам представления различают математические и визуальные (графические) структурные модели, причем многие математические модели могут быть представлены визуально.

Матрица корреляций содержит исчерпывающую информацию о структуре связей между переменными. Однако для ее непосредственного анализа требуются опыт и значительные усилия, особенно при большом количестве используемых переменных. Поэтому во многих случаях при анализе матрицы корреляций применяются специальные алгоритмы, реализованные в компьютерных программах, — метод главных компонент, факторный анализ, кластерный анализ и др.

Наиболее простым инструментом анализа матрицы корреляций является граф. *Граф* — это геометрическая схема, состоящая из точек, соединенных линиями и/или стрелками. Точки называются *вершинами* графа, линии — *ребрами*, стрелки — *дугами*. Вершины графа изображают переменные. При использовании графа для анализа корреляционных связей вершины соединяются линиями (ребрами). Если связь между переменными является причинной, вершины соединяются стрелками (дугами), направленными от причины к следствию. В графе одновременно могут присутствовать как ребра, так и дуги. Заметим, что если значения коэффициентов корреляции, по которым мы судим о силе связей и принимаем решение об отображении их на графе, *вычисляются*, то наличие или отсутствие причинных отношений между переменными, от которых зависит выбор дуги или ребра в качестве средства изображения связи на графе, *устанавливаются теоретически*. Подробнее эта тема обсуждается в разд. 1.2.

Вершины, соединенные ребрами и/или дугами, называются смежными. Смежность вершины равна числу проходящих через нее ребер и дуг. Вершины, не соединенные ни с какими другими вершинами (смежность равна 0), называются изолированными. Заметим, что независимо от количества изображенных связей на графе должны присутствовать *все* переменные (вершины), в том числе изолированные.

Наиболее сложной проблемой при построении графа связей является определение границы значений коэффициента корреляции, выше которой связь можно считать «существенной». Граница выбирается таким образом, чтобы граф, с одной стороны, был достаточно информативным, обеспечивающим отражение всех существенных, с точки зрения выявления структуры, связей между переменными, с другой стороны, был достаточно «прозрачным», не содержал избыточных связей, затрудняющих понимание структуры. В сложных случаях при построении графа рекомендуется для начала взять «завышенное» граничное значение, что обеспечит изображение минимального набора ребер. Впоследствии оно может быть уменьшено, что приведет к увеличению числа отображаемых связей.

Для построения графа рекомендуется начать с вершины с максимальной смежностью и изобразить все ее существенные связи. Затем перейти к следующей по смежности вершине. Процесс продолжается до тех пор, пока на граф не будут нанесены все ребра и/или дуги, а также изолированные вершины.

Взаимное расположение вершин на графе, обеспечивающее наиболее ясное представление структуры связей между переменными, нередко оказывается сложной дизайнерской задачей. Для ее решения могут применяться разнообразные программные средства; некоторые из них будут рассмотрены в гл. 4.

Пример 1.1 (продолжение)

В нашем примере в качестве границы выбрано $|r| \geq 0{,}55$ (рис. 1.1). В матрице корреляций (см. табл. 1.2) коэффициенты, значения которых по абсолютной величине не ниже 0,55, выделены жирным шрифтом. Причинно-следственные отношения между переменными установлены посредством логического анализа.

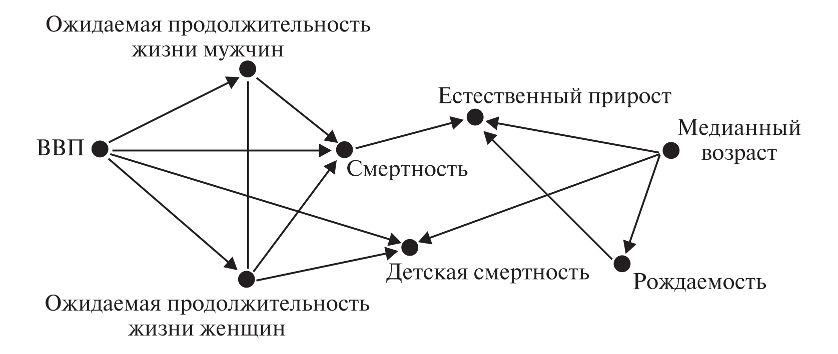


Рис. 1.1. Граф матрицы причинных связей между переменными

Две переменные – ВВП и медианный возраст, характеризующий возрастную структуру населения, – являются экзогенными; они не зависят от других рассматриваемых переменных, но в значительной мере сами их предопределяют. Уровень ВВП непосредственно влияет на ожидаемую продолжительность жизни мужчин и женщин, а также на уровень смертности (также зависящий от продолжительности жизни), и через него – опосредованно – на естественный прирост населения. С другой стороны, естественный прирост зависит от медианного возраста населения как непосредственно, так и опосредованно – через уровень рождаемости. Аналогично детская смертность зависит не только от возрастной структуры населения (страны со стареющим населением прилагают больше усилий для сохранения каждого ребенка), но также от экономического фактора (ВВП) как непосредственно, так и опосредованно – через увеличение продолжительности жизни женщин.

Самостоятельная работа

Постройте и проинтерпретируйте три графа корреляций между переменными, измеряющими успеваемость школьников по гуманитарным и математическим дисциплинам (табл. 1.3). Используйте последовательно границы $r \geq 0{,}45$; $r \geq 0{,}4$ и $r \geq 0{,}3$.

Какой из полученных графов наиболее информативен?

Таблица 1.3

Матрица корреляций: успеваемость по школьным дисциплинам

Переменная	x_1	x_2	x_3	x_4	x_5	x_6
x_1 – иностранный язык	1,00	0,53	0,51	0,29	0,33	0,25
x_2 – русский язык		1,00	0,45	0,40	0,32	0,33
x_3 – история			1,00	0,16	0,19	0,18
x_4 – арифметика				1,00	0,60	0,47
x_5 – алгебра					1,00	0,46
x_6 – геометрия						1,00

1.2. МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

Понятие причинной связи. Критерии каузальности. Причинной (причинно-следственной, каузальной) называется связь, в которой одни переменные интерпретируются как причины, другие – как следствия. Для обозначения причинных связей используются также понятия «зависимость», «влияние», «воздействие» и т. п. Переменная, измеряющая причину, называется независимой (предиктором); измеряющая следствие – зависимой[1].

[1] *Смелзер Н.* Социология. М., 1994. С. 14–16 ; 28–29.

В прикладной статистике причинно-следственная связь интерпретируется шире, чем в философии науки. Так, в философии связь рассматривается как причинная, только если наступление причины неизбежно влечет за собой наступление следствия и в отсутствие причины следствие не наступает, в статистике же оценивается *вероятность* наступления следствия в случае наличия причины (чем она выше, тем сильнее причинная связь).

В социально-экономических исследованиях существуют две традиции статистического изучения причинных связей. В рамках *социально-философской традиции*, восходящей к О. Конту, причинно-следственные отношения могут изучаться только посредством многофакторного эксперимента с обработкой полученных данных методами дисперсионного анализа[1]. *Эконометрическая традиция* предполагает применение моделей множественной регрессии. При регрессионном моделировании причинных отношений применяются ограничения на возможность использования переменных в качестве зависимых и независимых, предложенные М. Бором в 1949 г.[2]. Эти ограничения получили название критериев или принципов каузальности (причинности).

Первый критерий каузальности заключается в том, что причина должна по времени предшествовать следствию. Это ограничивает использование переменных, измеренных в ходе одного исследования. Например, по результатам одного опроса нельзя сказать, зависит ли достигнутый социальный статус респондента от имеющихся у него ценностей или ценности сформировались под влиянием достигнутого статуса. Исключение составляют показатели врожденного статуса — так называемые аскриптивные переменные (пол, возраст, среда происхождения, характеристики родительской семьи), — которые могут без ограничений использоваться в качестве независимых переменных. В некоторых случаях, при достаточном теоретическом обосновании, используются также показатели достигнутого статуса. Например, если зависимой переменной является заработная плата, в качестве независимых переменных могут выступать образование и должность, т. к. они достигнуты заведомо раньше, чем уровень зарплаты, декларируемый респондентом на момент исследования. Во всех остальных случаях причинные связи могут изучаться только экспериментально либо по данным панельных исследований, позволяющих измерить у одних и тех же респондентов независимые переменные на более ранних этапах жизненного пути, чем зависимые (например, достигнутая к определенному возрасту должность может зависеть от ценно-

[1] См.: *Гласс Дж., Стенли Дж.* Статистические методы в педагогике и психологии. М., 1976. С. 305–458.

[2] См.: *Sowa J. F.* Processes and Causality [Электронный ресурс]. URL: www.jfsowa.com.

стей респондента в период окончания средней школы). Следует, однако, иметь в виду, что более позднее событие или измерение не обязательно является следствием более раннего (*Post hoc, non ergo propter hoc!*[1]), т. е. причинный характер связи между переменными всегда нуждается в теоретическом обосновании.

Согласно *второму критерию*, между причиной и следствием должна существовать, как минимум, корреляционная связь. Действительно, если связи между двумя переменными нет, обсуждение того, является она причинной или корреляционной, теряет смысл.

Третий критерий состоит в том, что на взаимосвязь причины и следствия не должны влиять третьи факторы. Такое влияние может быть двух видов. Во-первых, оно может быть опосредующим, например, влияние пола на уровень доходов может опосредоваться занимаемой должностью. Во-вторых, третий фактор может обусловливать одновременно как причину, так и следствие. Например, социальный статус родителей может определять как образование респондента (независимая переменная), так и занимаемую должность (зависимая переменная). Данный критерий наиболее сложен для контроля, поскольку социологические переменные в большинстве случаев достаточно тесно коррелируют друг с другом. Общая рекомендация состоит в том, чтобы включить в модель *все* переменные, которые могут в изучаемой ситуации рассматриваться как независимые.

На рис. 1.2 представлены два вида влияния третьей переменной z на причинную связь между независимой переменной x и зависимой переменной y.

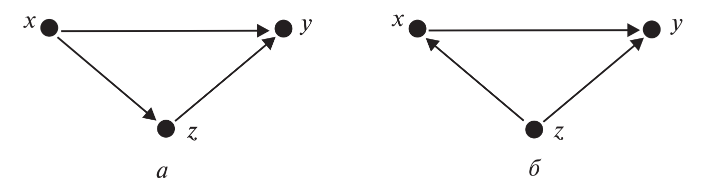


Рис. 1.2. Влияние фактора z на взаимодействие причины x и следствия y: *а* – опосредующее влияние; *б* – обусловливающее влияние

Необходимо отметить, что выполнение трех критериев каузальности является необходимым, но не достаточным условием причинной связи. Это означает, что, если хотя бы одно условие не соблюдается, связь не должна рассматриваться как причинная. Однако выполнение всех условий не гарантирует причинного характера связи, который может быть обоснован только теоретически.

[1] После этого, но не вследствие этого (лат.).

Задачи регрессионного анализа. Изучение причинно-следственных связей с использованием регрессионных моделей позволяет решать две важные исследовательские задачи – *объяснения* изменений зависимой переменной под воздействием одной или нескольких независимых переменных и *прогнозирования* наиболее вероятного значения зависимой переменной, если значения предикторов известны.

Виды уравнений регрессии. Регрессионные причинные модели используют только одну зависимую переменную[1]. Независимых переменных может быть одна (парная регрессия) или несколько (множественная регрессия). Далее мы будем рассматривать множественные модели как более общие. Во множественных моделях зависимую переменную принято обозначать y, независимые переменные (предикторы) – $x_1, x_2 \ldots x_k$, где k – количество независимых переменных. Зависимая переменная рассматривается как функция предикторов: $y = f(x_1, x_2 \ldots x_k)$.

Наиболее простой и часто используемой является модель множественной *линейной* регрессии, в которой функция представляет собой линейную комбинацию независимых переменных:

$$y = \sum_{i=1}^{k} b_i x_i + b_0 \, , \qquad (1.2)$$

где коэффициенты b_i ($i = \overline{0,k}$), называемые параметрами регрессии, используются как для объяснения изменчивости зависимой переменной y, так и для предсказания ее значения при заданных значениях независимых переменных x_i ($i = \overline{1,k}$).

Линейная регрессия первоначально была разработана для количественных нормально распределенных переменных. Однако в настоящее время в качестве независимых допускается использование дихотомических переменных и некоторых видов порядковых переменных, которые могут рассматриваться как квазиинтервальные[2]. Категориальные переменные (номинальные и порядковые, которые нельзя отнести к квазиинтервальным) могут быть включены в уравнение регрессии только посредством специальной техники фиктивных (*dummy*) переменных, которая будет рассмотрена ниже. В линейном регрессионном уравнении зависимая переменная может быть только количественной (квазиколичественной) или дихотомической.

Для зависимых категориальных переменных разработан целый ряд регрессионных моделей: дихотомическая логистическая регрессия, мультиноминальная логистическая регрессия, порядковая регрессия, пробит-регрессия и др. В данном пособии наряду с моделью линейной регрессии мы рассмотрим также логистическую модель.

---

[1] Наиболее простая модель с несколькими зависимыми переменными – путевой анализ – рассматривается в разд. 1.4.

[2] Например, переменные, измеренные по шкале Лайкерта.

Проблема мультиколлинеарности (интеркорреляции) тесно связана с проблемой выполнения третьего критерия каузальности. Мультиколлинеарностью называется сильная корреляционная связь между независимыми переменными, входящими в уравнение регрессии. Корреляция между независимыми переменными искажает показатели влияния каждой из них на зависимую переменную. Покажем этот эффект с помощью диаграмм Эйлера, применяемых в логике для демонстрации соотношения объемов и содержания понятий. В данном случае каждый круг обозначает полную дисперсию одной из переменных, условно принятую за единицу, пересечение кругов – коэффициент детерминации r^2 (квадрат коэффициента линейной корреляции), интерпретируемый как доля дисперсии одной переменной, объясненная взаимодействием со второй переменной (рис. 1.3).

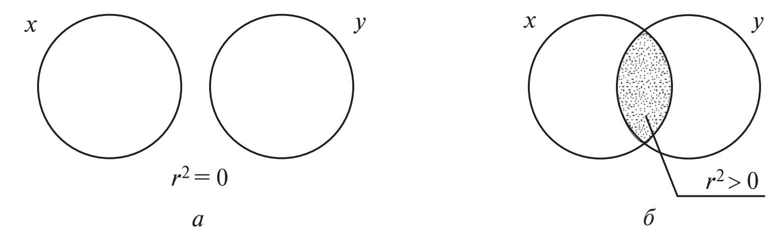


Рис. 1.3. Коэффициент детерминации для независимых переменных (*а*), и коррелирующих переменных (*б*)

Диаграммы Эйлера для простейшего случая множественной регрессии с двумя независимыми переменными представлены на рис. 1.4. Обе независимые переменные x_1 и x_2 коррелируют с зависимой переменной, но могут коррелировать или не коррелировать между собой. На рис. 1.4, *а* представлена модель множественной регрессии с некоррелирующими друг с другом независимыми переменными, на рис. 1.4, *б* – с коррелирующими независимыми переменными.

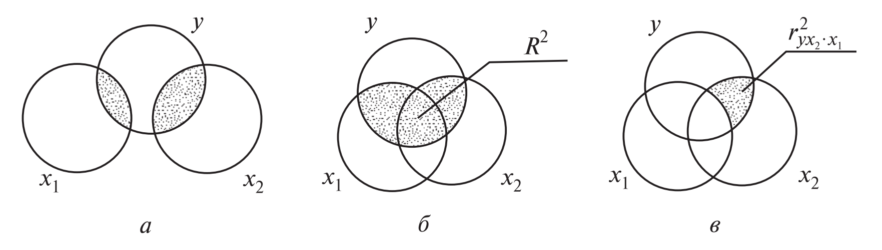


Рис.1.4. Взаимодействие трех переменных в уравнении регрессии: *а* – независимые переменные не коррелируют друг с другом; *б* – коэффициент множественной корреляции; *в* – частный коэффициент корреляции

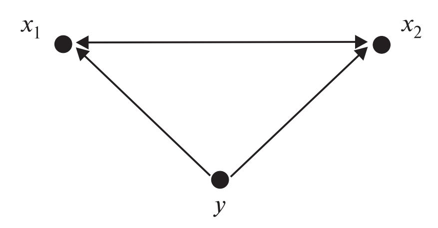


Рис. 1.5. «Ложная корреляция» между переменными x_1 и x_2, обусловленная влиянием переменной y

Для измерения совместного влияния независимых переменных на зависимую используется квадрат коэффициента множественной корреляции R^2 (R-квадрат). Если независимые переменные не коррелируют друг с другом, R^2 может быть получен как сумма квадратов коэффициентов линейной корреляции. Например, в случае двух некоррелирующих независимых переменных $R^2 = r^2_{x_1,y} + r^2_{x_2,y}$ (см. рис. 1.4, *а*).

Если независимые переменные коррелируют друг с другом, то R^2 не может быть вычислен просто как сумма двух или более коэффициентов детерминации (см. рис. 1.4, *б*), т. к. это приведет к завышенному результату. Для корректной его оценки применяется алгоритм, основанный на коэффициентах частной корреляции (см. рис. 1.4, *в*).

Заметим, что коэффициент частной корреляции может применяться не только при регрессионном моделировании, но также в случае так называемой «ложной корреляции», когда корреляционная связь между двумя переменными порождается одновременным влиянием на них третьей переменной (см. рис. 1.5).

Коэффициент частной корреляции служит для измерения связи между двумя переменными после устранения влияния одной или нескольких переменных, называемых контролируемыми. В зависимости от количества контролируемых переменных определяется порядок коэффициента частной корреляции – первый (одна переменная), второй (две переменные) и т. п. Применительно к *регрессионной модели* с k независимыми переменными максимальный порядок частного коэффициента корреляции составляет $k - 1$ (коэффициент корреляции между зависимой и одной из независимых переменных после устранения влияния остальных $k - 1$ переменных). Для набора k переменных, связанных друг с другом *корреляционными связями*, максимальный порядок коэффициента частной корреляции составляет $k - 2$.

Для обозначения частных коэффициентов корреляции используется сложный индекс: имена двух переменных, коэффициент между которыми измеряется, отделяется точкой от списка контролируемых переменных. Например, $r_{yx_1.x_2}$ – коэффициент 1-го порядка между переменными y и x_1 после устранения влияния контролируемой переменной x_2; $r_{yx_1.x_2x_3x_4}$ – коэффициент 3-го порядка между переменными y и x_1 после устранения влияния контролируемых переменных x_2, x_3 и x_4.

Коэффициент частной корреляции любого порядка можно вычислить по рекурсивной[1] формуле на основе коэффициентов предыдущего порядка. При вычислении коэффициента 1-го порядка используются обычные коэффициенты парной корреляции Пирсона, которые можно считать коэффициентами нулевого порядка:

$$r_{yx_1.x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_2}^2\right)\left(1 - r_{x_1x_2}^2\right)}}. \tag{1.3}$$

Формула коэффициента второго порядка выглядит следующим образом:

$$r_{yx_1.x_2x_3} = \frac{r_{yx_1.x_2} - r_{yx_3.x_2} r_{x_1x_3.x_2}}{\sqrt{\left(1 - r_{yx_3.x_2}^2\right)\left(1 - r_{x_1x_3.x_2}^2\right)}}. \tag{1.4}$$

Для ее применения необходимо предварительно вычислить три коэффициента 1-го порядка: $r_{yx_1.x_3}$, $r_{yx_2.x_3}$, $r_{x_1x_2.x_3}$. Этот процесс может быть продолжен вплоть до получения коэффициента частной корреляции порядка $k-1$:

$$r_{yx_1.x_2x_3\ldots x_k} = \frac{r_{yx_1.x_2x_3\ldots x_{k-1}} - r_{yx_k.x_2x_3\ldots x_{k-1}} r_{x_1x_k.x_2x_3\ldots x_{k-1}}}{\sqrt{\left(1 - r_{yx_k.x_2x_3\ldots x_{k-1}}^2\right)\left(1 - r_{x_1x_k.x_2x_3\ldots x_{k-1}}^2\right)}}. \tag{1.5}$$

Коэффициенты частной корреляции, как и парные коэффициенты, изменяются в пределах от –1 до +1 и являются симметричными, что позволяет представить их в виде треугольной матрицы.

Пример 1.1 (продолжение)

В табл. 1.4 представлена матрица корреляций между переменными, измеренными для некоторых европейских стран. Данная таблица получена из табл. 1.2 после удаления из списка переменных переменной «естественный прирост», т. к. она является линейной комбинацией переменных «рождаемость» и «смертность», что автоматически приводит к проблеме мультиколлинеарности.

Граф на рис. 1.1 позволяет предположить, что экзогенные переменные «медианный возраст» и «ВВП» влияют на взаимодействие всех остальных переменных. Устраним последовательно их влияние (табл. 1.5 и 1.6).

Сравнение коэффициентов парной и частной корреляции показывает, что эффект мультиколлинеарности может существенно искажать значения коэффициентов корреляции. Например, в матрице парных корреляций (табл. 1.4) коэффициент между рождаемостью и ВВП равен 0,08, что говорит практически об отсутствии связи. Однако коэффициент частной корреляции между теми же

[1] Рекурсивность означает, что для всех порядков формула сохраняет тот же вид, но изменяется порядок используемых в ней коэффициентов: для вычисления коэффициента частной корреляции порядка k используются коэффициенты порядка $(k-1)$.

Таблица 1.4

Матрица коэффициентов парной корреляции

Переменная	1	2	3	4	5	6	7
1. Медианный возраст	1	-0,74	0,26	-0,78	0,39	0,53	0,39
2. Рождаемость		1	-0,46	0,50	-0,01	-0,08	0,08
3. Смертность			1	-0,04	-0,75	-0,63	-0,55
4. Детская смертность				1	-0,49	-0,63	-0,58
5. Ожидаемая продолжительность жизни мужчин					1	0,92	0,77
6. Ожидаемая продолжительность жизни женщин						1	0,83
7. ВВП							1

Таблица 1.5

Матрица коэффициентов частной корреляции 1-го порядка (контролируемая переменная «медианный возраст»)

Переменная	2	3	4	5	6	7
2. Рождаемость	1	-0,41	-0,16	0,45	0,54	0,59
3. Смертность		1	0,27	-0,96	-0,94	-0,72
4. Детская смертность			1	-0,32	-0,41	-0,48
5. Ожидаемая продолжительность жизни мужчин				1	0,91	0,73
6. Ожидаемая продолжительность жизни женщин					1	0,80
7. ВВП						1

Таблица 1.6

Матрица коэффициентов парной корреляции 2-го порядка (контролируемые переменные «медианный возраст» и «ВВП»)

Переменная	2	3	4	5	6
2. Рождаемость	1	0,03	0,17	0,04	0,13
3. Смертность		1	-0,12	-0,91	-0,87
4. Детская смертность			1	0,05	-0,05
5. Ожидаемая продолжительность жизни мужчин				1	0,81
6. Ожидаемая продолжительность жизни женщин					1

переменными после устранения влияния переменной «медианный возраст» (см. табл. 1.5) составляет 0,59, следовательно, в странах с одинаковым медианным возрастом (одинаковой возрастной структурой) уровень рождаемости коррелирует с ВВП. С другой стороны, трудно объяснимый коэффициент корреляции между рождаемостью и смертностью (–0,46) после устранения влияния медианного возраста и ВВП (см. табл. 1.6) становится равным нулю (0,03).

Коэффициент множественной корреляции предназначен для измерения тесноты связи между зависимой переменной и всеми включенными в анализ предикторами (независимыми переменными). Для его обозначения также используется составной индекс $R_{y.x_1x_2...x_k}$: после имени зависимой переменной y ставится точка, и затем перечисляются имена независимых переменных x_1, x_2 ... x_k. Вычисление коэффициента множественной корреляции базируется на использовании последовательно вычисленных частных коэффициентов:

$$R_{y.x_1x_2...x_k} = \sqrt{1-\left(1-r^2_{yx_1}\right)\left(1-r^2_{yx_2.x_1}\right)\left(1-r^2_{yx_3.x_1x_2}\right)...\left(1-r^2_{yx_k.x_1x_2...x_{k-1}}\right)}\ . \qquad (1.6)$$

Значение коэффициента множественной корреляции изменяется в пределах от 0 до 1. Ниже будет показано, как он используется для оценки качества регрессионных моделей.

Этапы регрессионного анализа. Регрессионный анализ включает спецификацию модели, ее построение, интерпретацию, проверку гипотез о характере влияния независимых переменных на зависимую, прогнозирование значения зависимой переменной, оценку качества модели.

Спецификация модели (уравнения) регрессии включает выбор модели, определение зависимой и независимых переменных, подготовку при необходимости набора фиктивных (*dummy*) переменных.

Выбор регрессионной модели зависит от уровня измерения зависимой и независимой переменных. В частности, количественная зависимая переменная позволяет использовать модель множественной линейной регрессии. В свою очередь, выбор переменных определяется набором гипотез о факторах, влияющих на изменения зависимой переменной, и о характере влияния каждого из них.

Пример 1.2. Заработная плата сотрудников фирмы

Основная гипотеза: начальная зарплата сотрудника в организации зависит от образования, опыта работы, должности и пола.

Рабочие гипотезы:

1. Чем выше уровень образования, тем выше зарплата.
2. Чем больше опыт работы (стаж), тем выше зарплата.
3. Зарплата зависит от должности.
4. Зарплата женщин в среднем ниже, чем зарплата мужчин.

Уровень измерения переменных:

- зависимая переменная «зарплата», независимые переменные «образование» (в годах) и «стаж» (в месяцах) – количественные;
- независимые переменные «пол» (две градации) и «должность» (три градации – служащий, охранник, менеджер) являются номинальными и должны быть предварительно преобразованы.

Выбор модели: множественная линейная регрессия.

Оценка параметров уравнения множественной линейной регрессии. Уравнение множественной линейной регрессии существует в двух формах – нестандартизированной и стандартизированной. В *нестандартизированном уравнении* (формула 1.2) зависимая и независимые переменные представлены в оригинальном виде. *Стандартизированное уравнение* имеет вид

$$z_y = \sum_{i=1}^{k} \beta_i z_i \ , \tag{1.7}$$

где z_y – стандартизированная[1] переменная y; z_i ($i = \overline{1,k}$) – стандартизированные переменные x_i; β_i($i = \overline{1,k}$) – стандартизированные коэффициенты регрессии.

Для получения стандартизированных коэффициентов β_i чаще всего применяется метод наименьших квадратов: первые производные по всем параметрам модели регрессии приравнивают к нулю и получают систему из k уравнений:

$$\begin{aligned}
&\beta_1 + r_{1,2}\beta_2 + r_{1,3}\beta_3 + \ldots + r_{1,k}\beta_k = r_{1,y};\\
&r_{2,1}\beta_1 + \beta_2 + r_{2,3}\beta_3 + \ldots + r_{2,k}\beta_k = r_{2,y};\\
&r_{3,1}\beta_1 + r_{3,2}\beta_2 + \beta_3 + \ldots + r_{3,k}\beta_k = r_{3,y};\\
&\qquad\qquad\ldots\\
&r_{k,1}\beta_1 + r_{k,2}\beta_2 + r_{k,3}\beta_3 + \ldots + \beta_k = r_{k,y},
\end{aligned} \tag{1.8}$$

которую решают относительно параметров $\beta_1 \ldots \beta_k$. Коэффициенты корреляций между исходными переменными $r_{i,j}$ и $r_{i,y}$ ($i, j = \overline{1,k}$) вычисляются по исходным данным (формула 1.1).

Нестандартизированные коэффициенты вычисляются по формуле

$$b_i = \beta_i s_y / s_i \ \ (i = \overline{1,k}), \tag{1.9}$$

где s_y и s_i – стандартные отклонения переменных y и x_i соответственно.

Коэффициент b_0 (свободный член уравнения регрессии) вычисляется по формуле

$$b_0 = \overline{y} - \sum_{i=1}^{k} b_i \overline{x}_i \ . \tag{1.10}$$

[1] Стандартизированная переменная (z-оценка) для переменной x вычисляется по формуле $z = (x - \overline{x}) / s_x$, где $\overline{x}$ – среднее арифметическое; s_x – стандартное отклонение переменной x. Среднее арифметическое стандартизированной переменной $\overline{z} = 0$, стандартное отклонение $s_z = 1$.

Пример 1.2 (продолжение)

В табл. 1.7 представлена матрица корреляций (в форме нижнего треугольника), средние арифметические и стандартные отклонения для трех количественных переменных (зависимой и двух независимых).

Таблица 1.7

Матрица корреляций, средние арифметические и стандартные отклонения

Переменная	Зарплата	Образование	Стаж
Начальная зарплата (y)	1,00		
Образование (x_1)	0,63	1,00	
Стаж (x_2)	0,05	-0,25	1,00
Среднее арифметическое ($\bar{x}_i$)	17 016	13,5	96
Стандартное отклонение (s_i)	7 870	2,9	105

Система уравнений для определения стандартизированных коэффициентов β_i:

$$\beta_1 - 0{,}25\beta_2 = 0{,}63;$$
$$-0{,}25\beta_1 + \beta_2 = 0{,}05.$$

Стандартизированные коэффициенты (решение системы уравнений):

$$\beta_1 = 0{,}69;\ \beta_2 = 0{,}21.$$

Нестандартизированные коэффициенты:

$$b_1 = 0{,}69 \times 7870 / 2{,}9 = 1870;$$
$$b_2 = 0{,}21 \times 7870 / 105 = 16;$$
$$b_0 = 17016 - (1870 \times 13{,}5 + 16 \times 96) = -9765.$$

Таким образом, *уравнение регрессии* имеет следующий вид:

$$y = 1870x_1 + 16x_2 - 9765.$$

Интерпретация параметров уравнения регрессии. Каждый из нестандартизированных коэффициентов регрессии b_i показывает, на какую величину в среднем изменится предсказанное значение зависимой переменной y при увеличении значения соответствующей независимой переменной x_i на 1. Если $b_i > 0$, значение y увеличится, если $b_i < 0$, значение y уменьшится на величину $|b_i|$. Таким образом, знак коэффициента b_i (а также стандартизированного коэффициента β_i) соответствует направлению связи – прямой или обратной.

Свободный член нестандартизированного уравнения регрессии b_0 равен значению зависимой переменной y в случае, когда все независимые переменные $x_i = 0$. Значение b_0 содержательно интерпретируется, только если значение 0 входит в область определения каждого из предикторов x_i.

Стандартизированный коэффициент регрессии β_i используется для оценки силы влияния независимой переменной на зависимую. Напри-

мер, если $|\beta_1| > |\beta_2|$, то предиктор x_1 оказывает большее влияние на зависимую переменную y, чем предиктор x_2. Нестандартизированные коэффициенты b_i таким свойством не обладают, т. к. их значения зависят не только от силы связи, но и от масштаба измерения переменных.

Большинство современных программных средств представляют регрессионную модель в виде таблицы коэффициентов регрессии (табл. 1.8). Здесь и далее мы будем использовать именно такую форму представления регрессионных моделей. Имя зависимой переменной указывается под таблицей слева. Независимым переменным (их имена перечислены в первом столбце) соответствуют строки таблицы, причем первая строка («константа») соответствует свободному члену уравнения регрессии. Во втором столбце b_i находятся нестандартизированные коэффициенты регрессии: в первой строке значение b_0, далее значения b_i, соответствующие независимым переменным. В четвертом столбце (β_i) находятся значения стандартизированных коэффициентов для независимых переменных.

Пример 1.2 (продолжение)

Интерпретация параметров регрессии.

Стандартизированные коэффициенты имеют значения $\beta_1 = 0{,}69$, $\beta_2 = 0{,}22$. В данном случае $|\beta_1| > |\beta_2|$ ($\beta_1 = 0{,}69 > \beta_2 = 0{,}21$), следовательно, образование больше влияет на начальную зарплату в фирме, чем предшествующий опыт работы. Оба коэффициента положительные, значит, повышение уровня образования и увеличение стажа предыдущей работы повышают начальную зарплату при приеме на работу.

Нестандартизированные коэффициенты:

$b_1 = 1870$ – при повышении продолжительности образования на 1 год средняя начальная зарплата увеличивается на 1870 условных единиц.

$b_2 = 16$ – при увеличении стажа предшествующей работы на 1 месяц начальная зарплата в фирме увеличится в среднем на 16 условных единиц.

b_0 не интерпретируется, т. к. образование не может принять значение 0.

Таблица 1.8

Коэффициенты уравнения регрессии

Независимая переменная	Нестандартизированный коэффициент		Стандартизированный коэффициент	t	p
	b_i	$SE(b_i)$	β_i		
1	2	3	4	5	6
Константа	-9765	1417	–	-6,99	0,000
Образование (годы)	1870	97	0,69	19,42	0,000
Стаж (месяцы)	16	2,7	0,21	6,17	0,000

Зависимая переменная: «начальная зарплата».

Проверка гипотез о влиянии независимых переменных на зависимую в рамках регрессионной модели сводится к проверке знака и статистической значимости коэффициентов регрессии[1]. Знак коэффициента b_i характеризует *направленность* причинной связи между независимой переменной x_i и зависимой переменной y – является она прямой ($b_i > 0$) или обратной ($b_i < 0$). *Статистическая значимость* коэффициента регрессии заключается в том, что его значение для генеральной совокупности отличается от 0 с заданной доверительной вероятностью $1 - \alpha$, значения которой традиционно выбираются из чисел 0,99; 0,95; 0,9. Величина α, которая, соответственно, может принимать значения 0,01; 0,05; 0,1, называется уровнем значимости.

Для проверки гипотезы о статистической значимости коэффициента регрессии имеются две возможности. Первая возможность состоит в построении доверительного интервала[2] для коэффициента b_i с использованием значения его стандартной ошибки $SE(b_i)$ из 3-го столбца табл. 1.8. Если доверительный интервал, представленный на действительной оси, не включает значение 0, коэффициент регрессии является статистически значимым с соответствующей доверительной вероятностью. Если включает, коэффициент регрессии статистически значимым не является.

Пример 1.2 (продолжение)

При большом объеме выборки стандартная ошибка коэффициента регрессии имеет стандартное нормальное распределение. Поэтому при доверительной вероятности $1 - \alpha = 0,95$ доверительный коэффициент $Z_{0,975} = 1,96$, и 95 %-й доверительный интервал для коэффициента регрессии при независимой переменной «образование» составляет: (1870–1,96 × 97; 1870 + 1,96 × 97) или (1680; 2060). Значение «0» не входит в полученный интервал, соответственно коэффициент регрессии является статистически значимым при $\alpha = 0,05$. Кроме того, коэффициент регрессии имеет положительное значение. Следовательно, гипотеза о влиянии образования на начальную зарплату (с. 25) подтверждается.

Вторая возможность проверки статистической значимости коэффициента регрессии заключается в использовании t-критерия Стьюдента и связанного с ним p-значения[3] (столбцы 5, 6 в табл. 1.8). Процедура проверки заключается в сравнении p-значения из последнего столбца табл. 1.8 с выбранным уровнем статистической значимости α: если $p < \alpha$, коэффициент b_i является статистически значимым.

---

[1] Более строго статистическую гипотезу можно сформулировать математически: $H_0 : b_i = 0$; $H_1 : b_i > 0$ для прямой связи; $H_0 : b_i = 0$; $H_1 : b_i < 0$ для обратной связи.

[2] Подробнее см.: *Терещенко О. В.* Прикладная статистика для социальных наук. Минск, 2002. С. 56.

[3] В некоторых программных средствах (например, в SPSS) p-значение обозначается как *Sig* (от англ. *significance* – значимость).

Пример 1.2 (продолжение)

В нашем примере все коэффициенты регрессии являются статистически значимыми, т. к. соответствующие им $p < 0{,}000$, что заведомо меньше любого уровня значимости α. Кроме того, оба коэффициента регрессии являются положительными, что соответствует обсуждаемым гипотезам (с. 25).

Прогнозирование по уравнению регрессии. Уравнение регрессии позволяет предсказать среднее значение зависимой переменной при конкретной комбинации значений независимых переменных. Для этого нужно подставить соответствующие значения в уравнение регрессии:

$$\hat{y}(l)=\sum_{i=1}^{k} b_i x_i(l)+b_0, \tag{1.11}$$

где $x_i(l)$ – значение независимой переменной x_i ($i=\overline{1,k}$) для объекта l ($l=\overline{1,n}$); $\hat{y}(l)$ – предсказанное значение зависимой переменной y для объекта l.

Пример 1.2 (продолжение)

Предскажем значение зависимой переменной «начальная зарплата» ($\hat{y}$) для сотрудника с образованием 12 лет ($x_1 = 12$) и предшествующим стажем работы 20 месяцев ($x_2 = 20$): $\hat{y} = 1870 \times 12 + 16 \times 20 - 9765 = 12995$. Это значение интерпретируется как средняя начальная зарплата для сотрудников с образованием 12 лет и стажем работы 20 месяцев.

Оценка качества модели множественной регрессии. Аналогично тому, как это делается при оценке качества уравнения парной линейной регрессии, в уравнении множественной линейной регрессии полная дисперсия зависимой переменной s_y^2 может быть представлена как сумма двух составляющих – дисперсии, *объясненной* влиянием набора независимых переменных $x_1, x_2 \dots x_k$ ($s_{об}^2$), и *остаточной* дисперсии ($s_{ост}^2$), порожденной неучтенными факторами: $s_y^2 = s_{об}^2 + s_{ост}^2$. Показателем качества (объясняющей способности) модели является доля объясненной дисперсии в общей дисперсии зависимой переменной. Можно показать, что она численно равна квадрату коэффициента множественной корреляции (см. формулу 1.6):

$$s_{об}^2/s_y^2 = \sum_{l=1}^{n}(\hat{y}(l)-\bar{y})^2 \Big/ \sum_{l=1}^{n}(y(l)-\bar{y})^2 = R^2, \tag{1.12}$$

где $y(l)$ – значение зависимой переменной y для объекта из выборки с номером l ($l=\overline{1,n}$); $\hat{y}(l)$ – предсказанное значение зависимой переменной для того же объекта; $\bar{y}$ – среднее арифметическое переменной y.

Пример 1.2 (продолжение)

Для данного примера коэффициент множественной корреляции $R = 0{,}668$, его квадрат $R^2 = 0{,}446$, следовательно, дисперсия независимой переменной «началь-

ная зарплата» на 44,6 % объясняется двумя независимыми переменными – образованием и стажем предшествующей работы.

Включая в уравнение поочередно дополнительные независимые переменные, можно оценить, насколько при этом увеличивается значение R^2 в абсолютном исчислении и насколько значимо это увеличение относительно первоначального значения R^2. Таким образом можно оценить вклад каждого следующего предиктора в объяснение дисперсии зависимой переменной y. Однако необходимо иметь в виду, что вследствие эффекта мультиколлинеарности результаты могут существенно зависеть от порядка включения независимых переменных в модель.

Множественная регрессия с категориальными независимыми переменными. Как отмечалось выше, «классическая» модель множественной линейной регрессии предполагает использование количественных переменных с уровнем измерения не ниже интервального. Однако в социологии и смежных социальных науках количественные переменные встречаются достаточно редко, преобладают категориальные переменные, измеряемые по номинальным и порядковым шкалам. Здесь мы рассмотрим три вида таких переменных: дихотомические, номинальные с двумя градациями, номинальные и порядковые с числом градаций больше двух. В последних двух случаях применяется подход, получивший название *dummy*-кодирования[1] и позволяющий из любой категориальной переменной создать набор дихотомических переменных, каждая из которых фиксирует одно свойство объекта.

Дихотомические независимые переменные используются в регрессионных моделях практически без ограничений в своем обычном виде: 0 – отсутствие некоторого свойства у объекта (например, респондента); 1 – наличие. Коэффициент регрессии при независимой дихотомической переменной показывает, насколько в среднем изменится значение зависимой переменной y при наличии данного свойства по сравнению с его отсутствием, что соответствует общему случаю интерпретации коэффициента регрессии, т. к. «переход» от значения 0 к значению 1 соответствует «увеличению» значения дихотомической переменной на 1.

Номинальная независимая переменная из двух категорий может быть перекодирована в дихотомическую переменную. При этом значением 1 ко-

[1] Техника *dummy*-кодирования переменных (на русский язык нередко переводится как фиктивные переменные) заимствована социологией из эконометрики, где используется достаточно давно. Автор методики неизвестен, на что указывает видный экономист, председатель Международной экономической ассоциации в 1971–1974 гг. Фриц Махлуп (см.: *Machlup F.* Proxies and Dummies // Journal of Political Economy. 1974. № 82. P. 892).

дируется модальная (наиболее часто встречающаяся) категория, или категория, более важная с точки зрения проверки гипотез или анализа данных. Например, если проверяется гипотеза о том, что заработная плата женщин в среднем ниже, чем зарплата мужчин, женский пол следует кодировать значением 1, мужской – значением 0.

Пример 1.2 (продолжение)

Добавим к уравнению регрессии переменную «пол» (табл. 1.9), т. е. будем изучать влияние на зарплату трех независимых переменных: образование, стаж работы и пол. В соответствии с гипотезой (с. 25) пол закодирован дихотомической переменной со значениями: 1– женский, 0 – мужской.

Таблица 1.9

Коэффициенты уравнения множественной регрессии с дихотомической переменной «пол»

Независимая переменная	Нестандартизированный коэффициент		Стандартизированный коэффициент	*t*	*p*
	b	*SE*(*b*)	β		
Константа	-4492	1647	–	-2,73	0,007
Образование (годы)	1625	103	0,60	15,82	0,000
Стаж (месяцы)	12	2,7	0,16	4,47	0,000
Пол (женский)	-3447	583	-0,22	-5,91	0,000

Зависимая переменная: «начальная зарплата».

Коэффициент при независимой переменной «женский пол» равен –3447, из чего следует, что начальная зарплата женщин в среднем на 3447 условных единиц ниже, чем начальная зарплата мужчин, что подтверждает проверяемую гипотезу.

Отметим также, что введение в уравнение новой независимой переменной привело к изменению значений коэффициентов регрессии при предикторах «образование» и «стаж» (вследствие эффектов интеркорреляции), однако они остались статистически значимыми.

Значение R^2 увеличилось и составило 48,4 %. Таким образом, учет пола сотрудника повысил объясняющую способность модели на 3,8 %.

Категориальные переменные с числом градаций больше двух также могут быть включены в уравнение регрессии посредством *dummy*-кодирования. Оно заключается в том, что номинальная переменная с k градациями преобразуется в $k - 1$ дихотомическую переменную, причем потери исходной информации не происходит.

Использование *dummy*-кодирования предполагает объявление одного из значений референтным (базовым, ссылочным) и рассмотрение всех остальных значений в сравнении с ним. В качестве референтного может выбираться модальное значение номинальной переменной или, напротив, значение, представляющее наименьший интерес. Для порядковой пере-

менной референтным может стать значение, стоящее в центре распределения, а также максимальное или минимальное значение.

Дихотомические переменные создаются для всех значений, кроме референтного. Они называются *dummy*-переменными или фиктивными переменными. Для каждого респондента в наборе *dummy*-переменных только одна может принимать значение 1, та, которая соответствует категории, выбранной респондентом. Если респондент выбирает референтное значение, все *dummy*-переменные равны.

Интерпретация коэффициентов регрессии для *dummy*-переменных осуществляется по отношению к референтной категории: коэффициент показывает, насколько в среднем изменится значение зависимой переменной *y* для категории, представленной *dummy*-переменной, по сравнению с референтной категорией.

Пример 1.2 (продолжение)

Согласно одной из гипотез (с. 25), начальная зарплата зависит от должности, на которую претендует соискатель. Допустим для простоты, что в организации три основных вида должностей: служащие, охранники и менеджеры, причем бо́льшая часть всех работников является служащими. Выберем эту категорию в качестве референтной, а для двух других создадим *dummy*-переменные «охранник» и «менеджер». Схема кодирования представлена в табл. 1.10.

Таблица 1.10

Схема *dummy*-кодирования переменной «должность»

Должность	Исходный код категории	Коды *dummy*-переменных	
		«охранник»	«менеджер»
Служащий (референтная)	1	0	0
Охранник	2	1	0
Менеджер	3	0	1

Для проверки гипотезы о влиянии должности на начальную зарплату введем в уравнение регрессии две *dummy*-переменные «охранник» и «менеджер» вместо исходной переменной «должность», референтное значение указано в скобках (табл. 1.11). Прежде всего отметим, что введение «должности» существенно повысило качество регрессионной модели: ее объясняющая способность достигла 70 % ($R^2 = 0{,}7$).

Анализ нестандартизированных коэффициентов регрессии показывает, что должность менеджера позволяет получить значительно более высокую начальную зарплату, чем должность служащего (на 12134 условные единицы больше). В то же время разность средней начальной зарплаты охранника и служащего хоть и кажется довольно значительной (–999 условных единиц), не является статистически значимой ($p = 0{,}335$).

Анализ стандартизированных коэффициентов регрессии позволяет заключить, что наибольшее влияние на начальную зарплату оказывает должность менеджера ($\beta = 0{,}6$). В силу того, что должность существенно зависит от уровня образования, стандартизированный и нестандартизированный коэффициенты регрессии при «образовании» с введением в уравнение «должности» значительно уменьшились.

Таблица 1.11

Коэффициенты уравнения множественной регрессии с набором *dummy*-переменных «должность»

Независимая переменная	Нестандартизированный коэффициент		Стандартизированный коэффициент	t	p
	b	$SE(b)$	β		
Константа	6150	1426	–	4,31	0,000
Образование	665	97	0,24	6,87	0,000
Стаж	11	2,2	0,14	4,78	0,000
Пол женский	-2660	467	-0,17	-5,70	0,000
Должность (*служащий*):					
охранник	-999	1035	-0,03	-0,97	0,335
менеджер	12134	662	0,60	18,33	0,000

Зависимая переменная: «начальная зарплата».

Множественная регрессия с зависимыми категориальными переменными. Наиболее простым случаем в этом классе регрессионных моделей является множественная линейная модель с дихотомической зависимой переменной, фиксирующей наступление или ненаступление некоторого события. В этом случае предсказанное значение $\hat{y}(l)$ для респондента с номером l интерпретируется как вероятность того, что событие, закодированное значением 1, для данного респондента произойдет.

Предсказанные в рамках данной модели значения зависимой переменной $\hat{y}(l)$ в большинстве укладываются в интервал [0; +1], однако некоторые значения могут выходить за пределы этого интервала. В этом случае используют простую перекодировку: значения $\hat{y}(l) > 1$ перекодируются в $\hat{y}(l) = 1$; значения $\hat{y}(l) < 0$ перекодируются в $\hat{y}(l) = 0$. Соответственно, в первом случае прогнозируется вероятность наступления события, равная 100 %, во втором – вероятность, равная 0 %.

Пример 1.3. Поступление на дневную бюджетную форму образования после окончания среднего учебного заведения

Зависимая переменная – получение высшего образования (дихотомическая). Независимые переменные: тип среднего образования (номинальная), тип поселения (порядковая), пол (номинальная), успеваемость – средний балл по аттестату (количественная), образование отца (порядковая), образование матери (порядковая).

Переменная «пол» дихотомизирована: 1 – женский, 0 – мужской.

Остальные номинальные и порядковые переменные подвергнуты *dummy*-кодированию. Таким образом, в уравнении 13 независимых переменных, в том числе 1 дихотомическая, 1 количественная и 11 фиктивных.

Уравнение множественной регрессии представлено в табл. 1.12, объясняющая способность модели $R^2 = 0{,}318$. Референтные значения *dummy*-переменных указаны в скобках курсивом.

Таблица 1.12

Коэффициенты уравнения множественной линейной регрессии с дихотомической зависимой переменной

Независимая переменная	Нестандартизированный коэффициент		Стандартизированный коэффициент	t	p
	b	SE(b)	β		
Константа	-0,56	0,15	–	-3,69	0,000
Среднее образование (*общее среднее*):					
ПТУ	-0,32	0,06	-0,24	-5,67	0,000
колледж	-0,10	0,05	-0,09	-2,14	0,033
спецшкола	0,14	0,07	0,08	2,07	0,039
Тип поселения (*город*):					
село	-0,12	0,05	-0,09	-2,25	0,025
столица	0,08	0,05	0,07	1,73	0,084
Пол женский	0,06	0,04	0,06	1,64	0,101
Успеваемость	0,11	0,02	0,27	6,62	0,000
Образование отца (*общее среднее*):					
базовое сред.	-0,03	0,05	-0,03	-0,49	0,623
среднее спец.	-0,01	0,06	-0,01	-0,13	0,896
высшее	0,23	0,07	0,18	3,51	0,000
Образование матери (*общее среднее*):					
базовое сред.	0,02	0,06	0,02	0,40	0,688
среднее спец.	0,07	0,06	0,06	1,23	0,221
высшее	0,13	0,07	0,09	1,80	0,072

Зависимая переменная: «поступление на бюджет».

Предскажем вероятность продолжения образования для респондента мужского пола, со средним баллом по аттестату 8,5, закончившего техникум в городе, не являющемся столицей, отец которого имеет высшее образование, мать – среднее специальное (табл. 1.13). При вычислении предсказанного значения зависимой переменной (вероятности продолжения образования) для респондента с соответствующими характеристиками используются значения коэффициентов регрессии

и значения независимых переменных, в том числе *dummy*-переменных. В данном случае предсказанное значение зависимой переменной $\hat{y} = 0{,}58$ (вероятность продолжения образования 58 %) получено как алгебраическая сумма значений в последнем столбце табл. 1.13.

Таблица 1.13

Вычисление предсказанного значения зависимой переменной при использовании независимых *dummy*-переменных

Независимая переменная	Нестандартизированный коэффициент b_i	Значения независимых переменных x_i	$b_i x_i$
Константа	-0,56		-0,56
Среднее образование (*общее среднее*):			
ПТУ	-0,32	0	0
колледж	-0,10	1	-0,10
спецшкола	0,14	0	0
Тип поселения (*город*):			
село	-0,12	0	0
столица	0,08	0	0
Пол женский	0,06	0	0
Успеваемость	0,11	8,5	0,94
Образование отца (*общее среднее*):			
базовое сред.	-0,03	0	0
среднее спец.	-0,01	0	0
высшее	0,23	1	0,23
Образование матери (*общее среднее*):			
базовое сред.	0,02	0	0
среднее спец.	0,07	1	0,07
высшее	0,13	0	0

В более сложных случаях, когда число градаций зависимой переменной превышает два, применяется логистическая регрессия, базирующаяся на *dummy*-кодировании зависимой переменной. Для порядковых зависимых переменных применяется порядковая логистическая регрессия[1].

Последовательный отбор независимых переменных. Если выбранные для регрессионной модели независимые переменные тесно коррелируют друг с другом, их одновременное включение в модель нецелесообразно,

[1] *Наследов А. Д.* SPSS: Компьютерный анализ данных в психологии и социальных науках. СПб., 2007 ; *Бююль А., Цёфель П.* SPSS: Искусство обработки информации. М., 2002.

т. к. приводит к эффектам мультиколлинеарности. Существуют два подхода к решению этой проблемы. Первый подход – традиционный (теоретический) – заключается в том, чтобы из каждой группы коррелирующих между собой предикторов выбрать один, наиболее важный с точки зрения рассматриваемых теоретических гипотез. Второй подход – «технологический» – состоит в том, чтобы, используя быстродействие компьютеров, отобрать и включить в модель наиболее информативные независимые переменные. Эту функцию, реализованную во многих современных программных средствах, рекомендуется использовать только на начальном («разведочном») этапе построения модели или в тех редких случаях, когда потенциальные предикторы взаимозаменяемы с точки зрения проверяемых гипотез.

Критерием отбора переменных для включения в модель могут служить абсолютная величина частного коэффициента корреляции и увеличение квадрата коэффициента множественной корреляции R^2. Включение новых переменных в модель прекращается, когда увеличение R^2 перестает быть статистически значимым. Кроме увеличения R^2, показателем улучшения качества модели может служить также уменьшение стандартной ошибки предсказанного значения зависимой переменной.

Заметим, что при использовании категориальных независимых переменных в модель должны включаться дихотомические индикаторы, представляющие *все* категории, чтобы можно было отобрать наиболее информативные из них. Поэтому при последовательном отборе независимых переменных применяется не *dummy*-, а *индикаторное кодирование*, при котором количество дихотомических переменных-индикаторов совпадает с количеством категорий (табл. 1.14).

Отметим также, что последовательный отбор независимых переменных может осуществляться для всех видов логистической регрессии, которые мы рассмотрим ниже. Единственное отличие состоит в том, что независимые категориальные переменные включаются в логистические модели полностью и подвергаются *dummy*-кодированию непосредственно в процессе построения модели.

Пример 1.2 (продолжение)

В список независимых переменных включены должность, представленная тремя дихотомическими индикаторами (менеджер, охранник, служащий), пол и принадлежность к национальным меньшинствам (дихотомические переменные), образование и возраст (в годах), предшествующий стаж работы (в месяцах). Последовательный отбор наиболее информативных переменных позволил получить 4 «вложенные» модели. Последовательность включения переменных в модель представлена в табл. 1.15. Полученные результаты позволяют заключить, что основными факторами, влияющими на величину начальной зарплаты, являются должность менеджера, пол, образование и стаж работы.

Таблица 1.14

Схема индикаторного кодирования переменной «должность»

Должность	Исходный код категории	Код индикаторных переменных		
		«служащий»	«охранник»	«менеджер»
Служащий	1	1	0	0
Охранник	2	0	1	0
Менеджер	3	0	0	1

Таблица 1.15

Последовательность включения переменных в модель

№ модели	Независимая переменная	Нестандартизированный коэффициент		Стандартизированный коэффициент β	*t*	*p*
		b	*SE*(*b*)			
1	Константа	14148	249	–	56,92	0,000
	Менеджер	16109	591	0,78	27,26	0,000
2	Константа	16098	332	–	48,46	0,000
	Менеджер	14598	583	0,71	25,05	0,000
	Пол (женский)	-3682	447	-0,23	-8,24	0,000
3	Константа	8888	1248	–	7,12	0,000
	Менеджер	12339	677	0,60	18,21	0,000
	Пол (женский)	-3109	442	-0,20	-7,04	0,000
	Образование (годы)	545	91	0,20	5,98	0,000
4	Константа	5822	1379	–	4,22	0,000
	Менеджер	12156	663	0,59	18,33	0,000
	Пол (женский)	-2524	449	-0,16	-5,62	0,000
	Образование (годы)	685	94	0,25	7,30	0,000
	Стаж (месяцы)	10	2,1	0,13	4,77	0,000

Зависимая переменная: «начальная зарплата».

В табл. 1.16 представлены независимые переменные, не включенные в модель на каждом из этапов.

Таблица 1.16

Переменные, не включенные в модель

№ модели	Независимая переменная	β	*t*	*p*	Частная корреляция
1	Образование (годы)	0,25	7,32	0,000	0,320
	Служащий	-0,05	-1,02	0,309	-0,047
	Охранник	0,03	1,02	0,309	0,047
	Стаж (месяцы)	0,11	3,88	0,000	0,176
	Пол (женский)	-0,23	-8,24	0,000	-0,355
	Возраст	0,06	2,01	0,046	0,092
	Национальное меньшинство	-0,01	-0,21	0,831	-0,010
2	Образование (годы)	0,20	5,98	0,000	0,266
	Служащий	0,06	1,25	0,214	0,057
	Охранник	-0,04	-1,25	0,214	-0,057
	Стаж (месяцы)	0,07	2,48	0,014	0,114
	Возраст	0,06	2,36	0,019	0,108
	Национальное меньшинство	-0,04	-1,46	0,144	-0,067
3	Служащий	-0,05	-0,86	0,389	-0,040
	Охранник	0,03	0,86	0,389	0,040
	Стаж (месяцы)	0,13	4,77	0,000	0,215
	Возраст	0,12	4,45	0,000	0,201
	Национальное меньшинство	-0,03	-1,20	0,230	-0,055
4	Служащий	0,05	0,96	0,337	0,044
	Охранник	-0,03	-0,96	0,337	-0,044
	Возраст	0,05	1,06	0,292	0,049
	Национальное меньшинство	-0,04	-1,68	0,093	-0,078

Изменение квадрата множественной корреляции и стандартной ошибки предсказанных значений зависимой переменной для моделей из табл. 1.15 представлено в табл. 1.17.

Таблица 1.17

Изменение R^2 и стандартной ошибки предсказанных значений зависимой переменной

Модель	R	R^2	Стандартная ошибка предсказанных значений зависимой переменной
1	0,782	0,612	4912
2	0,813	0,661	4596
3	0,828	0,685	4435
4	0,836	0,700	4336

Самостоятельная работа

Проинтерпретируйте уравнение регрессии (см. табл. 1.11) и значение $R^2 = 0{,}318$.

1.3. ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

Логистическая регрессия предназначена для изучения причинных связей между категориальными переменными. Для построения уравнений логистической регрессии используется метод максимального правдоподобия. Зависимая переменная может быть только категориальной. Независимые переменные могут быть двух видов – категориальные и количественные («метрические»). Категориальные независимые переменные в уравнении логистической регрессии по традиции называют *факторами*, количественные – *ковариатами*. В современном статистическом программном обеспечении процедура *dummy*-кодирования категориальных переменных (зависимой и независимых) в логистической регрессии, как правило, осуществляется автоматически, в отличие от «классической» множественной линейной регрессии, где ее в большинстве случаев приходится осуществлять вручную.

Модели логистической регрессии различаются по виду зависимой переменной: бинарная (зависимая переменная с двумя градациями), мультиноминальная (зависимая переменная номинальная с числом градаций больше двух), порядковая (зависимая переменная порядковая).

Бинарная логистическая регрессия предполагает использование дихотомизированной зависимой переменной. Однако в отличие от множественной линейной регрессии с дихотомической зависимой переменной, рассмотренной в разд. 1.2, линейное уравнение используется здесь для оценки не собственно вероятности наступления события, а для оценки логарифма отношения вероятности наступления события к вероятности его ненаступления – *логита*. Данный подход происходит от логлинейного анализа

многомерных таблиц сопряженности[1], расширенного за счет возможности использовать количественные ковариаты.

Обозначим вероятность события, закодированного значением 1, буквой p. Вероятность ненаступления события будет равна $1 - p$. Отношение этих двух вероятностей $p/(1 - p)$ представляет собой отношение шансов наступления и ненаступления события.

Например, если $p = 0{,}5$, то $p/(1 - p) = 0{,}5 / (1 - 0{,}5) = 0{,}5 / 0{,}5 = 1$, т. е. шансы равны. Если $p = 0{,}25$, то $p/(1 - p) = 0{,}25 / (1 - 0{,}25) = 0{,}25 / 0{,}75 = 1/3$, т. е. наступление события в 3 раза менее вероятно, чем его ненаступление.

Натуральный логарифм отношения шансов $\ln\left[p/(1-p)\right]$ (логит) оценивается с помощью линейного уравнения:

$$\ln\left[\frac{p}{1-p}\right] = B_0 + \sum_{i=1}^{k} B_i x_i = B_0 + B_1 x_1 + B_2 x_2 + \ldots + B_k x_k, \qquad (1.13)$$

где $x_i (i = \overline{1,k})$ — независимые переменные, представляющие собой *dummy*-переменные, полученные из категориальных факторов, и/или количественные ковариаты.

Коэффициенты B_i интерпретируются аналогично коэффициентам множественной линейной регрессии, однако здесь они показывают, насколько в среднем изменится *логит* при изменении независимой переменной.

Этот способ представления результатов не очень удобен для интерпретации, поэтому формулу (1.13) принято представлять в преобразованном виде[2]:

$$\frac{p}{1-p} = e^{B_0 + B_1 x_1 + B_2 x_2 + \ldots + B_k x_k} = e^{B_0} e^{B_1 x_1} e^{B_2 x_2} \ldots e^{B_k x_k}. \qquad (1.14)$$

Интерпретации, как правило, подлежат значения e^{B_i}, каждое из которых показывает, как изменится отношение шансов $p/(1 - p)$ при увеличении значения независимой переменной x_i на единицу. Если $B_i > 0$, то $e^{B_i} > 1$ и отношение вероятностей возрастет в e^{B_i} раз. Если $B_i < 0$, то $e^{B_i} < 1$ и отношение вероятностей соответственно уменьшится. Если $B_i = 0$, то $e^{B_i} = 1$, т. е. независимая переменная x_i на отношение вероятностей не влияет.

Если необходимо предсказать не отношение шансов $p/(1 - p)$, а собственно вероятность наступления события p, уравнение логистической регрессии может быть представлено в следующем виде:

$$p = \frac{e^{(B_0 + B_1 x_1 + B_2 x_2 + \ldots + B_k x_k)}}{1 + e^{(B_0 + B_1 x_1 + B_2 x_2 + \ldots + B_k x_k)}}. \qquad (1.15)$$

---

[1] *Аптон Г.* Анализ таблиц сопряженности. М., 1982.

[2] В данном преобразовании используется соотношение натурального логарифма и экспоненты: если $\ln(x) = a$, то $x = e^a$.

Уравнение бинарной логистической регрессии в современном программном обеспечении принято представлять в виде таблицы.

Пример 1.3 (продолжение)

В табл. 1.18 представлено уравнение бинарной логистической регрессии, в котором зависимая переменная – логарифм отношения шансов продолжения и непродолжения образования. Значения коэффициентов B_i даются во втором столбце, e^{B_i} – в последнем столбце таблицы. Статистическая значимость коэффициентов B_i может быть проверена двумя способами: через построение доверительного интервала аналогично тому, как это описано на с. 29 (значения стандартных ошибок в третьем столбце табл. 1.18). Второй способ состоит в использовании статистики Уальда (*Wald*), которая позволяет определить статистическую значимость не только отдельных индикаторов, но и номинальных независимых переменных в целом. Значение статистики Уальда приведено в четвертом столбце, число степеней свободы – в пятом, *p*-значение – в шестом.

Таблица 1.18

Коэффициенты уравнения логистической регрессии

Независимая переменная	B_i	*SE*(*B*)	*Wald*	*df*	*p*	e^{B_i}
1	2	3	4	5	6	7
Среднее образование (*общее среднее*):	–	–	28,30	3	0,000	–
ПТУ	-2,53	0,56	20,41	1	0,000	0,08
колледж	-0,57	0,28	4,14	1	0,042	0,57
спецшкола	0,77	0,44	2,99	1	0,084	2,15
Тип поселения (*город*):	–	–	5,65	2	0,059	–
село	-0,45	0,35	1,70	1	0,193	0,64
столица	0,43	0,28	2,45	1	0,118	1,54
Пол женский	0,45	0,26	3,09	1	0,079	1,57
Успеваемость	0,71	0,12	34,01	1	0,000	2,03
Образование отца (*общее среднее*):	–	–	14,72	3	0,002	–
базовое сред.	0,01	0,39	0,00	1	0,972	1,01
среднее спец.	0,11	0,37	0,08	1	0,771	1,11
высшее	1,45	0,44	10,95	1	0,001	4,25
Образование матери (*общее среднее*):	–	–	1,49	3	0,684	–
базовое сред.	0,00	0,39	0,00	1	0,993	1,00
среднее спец.	0,26	0,36	0,52	1	0,471	1,29
высшее	0,43	0,43	0,98	1	0,323	1,53
Константа	-6,56	1,11	35,06	1	0,000	0,00

Зависимая переменная: «поступление на бюджет».

Анализ табл. 1.18 показывает, что увеличение логита (коэффициенты B_i) и отношения шансов в пользу продолжения образования (e^{B_i}) в первую очередь определяются такими характеристиками, как высшее образование отца, получение среднего образования в специальных школах и классах, высокая успеваемость. Наиболее низкие шансы получить высшее образование имеют выпускники ПТУ. В целом наибольший вклад в объяснение продолжения образования вносят такие факторы, как тип полученного среднего образования, успеваемость, образование отца.

Определим теперь вероятность продолжения образования для респондента мужского пола со средним баллом по аттестату 8,5, закончившего техникум в городе, не являющемся столицей, отец которого имеет высшее образование, мать – среднее специальное. Для этого вычислим $\ln\left[\frac{p}{1-p}\right] = B_0 + B_1x_1 + B_2x_2 + \ldots + B_kx_k = 0{,}59$ как алгебраическую сумму чисел в последнем столбце табл. 1.19. Тогда $\frac{p}{1-p} = e^{b_0+b_1x_1+b_2x_2+\ldots+b_kx_k} = e^{0{,}59} = 1{,}8$, откуда получаем $p = 0{,}64$, что в целом достаточно близко к результату, полученному методом множественной линейной регрессии для зависимой дихотомической переменной $\hat{y} = 0{,}58$ (с. 36).

Таблица 1.19

Вычисление логита

Независимая переменная	B_i	e^{B_i}	x_i	B_ix_i
Среднее образование (*общее среднее*): ПТУ	-2,53	0,08	0	
колледж	-0,57	0,57	1	-0,57
спецшкола	0,77	2,15	0	
Тип поселения (*город*): село	-0,45	0,64	0	
столица	0,43	1,54	0	
Пол женский	0,45	1,57	0	
Успеваемость	0,71	2,03	8,5	6,01
Образование отца (*общее среднее*): базовое сред.	0,01	1,01	0	
среднее спец.	0,12	1,11	0	
высшее	1,45	4,25	1	1,45

Окончание табл. 1.19

Независимая переменная	B_i	e^{B_i}	x_i	$B_i x_i$
Образование матери (*общее среднее*): базовое сред.	0,00	1,00	0	
среднее спец.	0,26	1,29	1	0,26
высшее	0,43	1,53	0	
Константа	-6,56	0,0014	–	-6,56

Заметим, что при использовании современного программного обеспечения не приходится вычислять вероятности наступления события «вручную», т. к. предсказанные вероятности для каждого респондента могут быть сохранены в качестве значений дополнительной переменной и в дальнейшем проанализированы.

Качество бинарной логистической модели оценивается двумя способами – с помощью мер соответствия модели исходным данным и по классификационной таблице. Аналогами квадрата коэффициента множественной корреляции (R^2) для бинарной логистической модели являются *псевдо*-R^2 Кокса и Шелла и *псевдо*-R^2 Нагелькерке. Последовательно добавляя в модель независимые переменные, можно оценить их полезность по увеличению данных мер связи, разработанных специально для логистической регрессии.

Классификационная таблица используется не только в логлинейной регрессии, но и во многих моделях, в частности в модели дискриминантного анализа. Она показывает процент правильной классификации в каждой из категорий зависимой переменной, а также по выборке в целом.

Пример 1.3 (продолжение)

Для уравнения, представленного в табл. 1.18, значения коэффициентов псевдо-R^2 Кокса и Шелла и псевдо-R^2 Нагелькерке составляют соответственно 0,324 и 0,436, из чего следует, что объяснительная способность модели достаточно высока. В модели с дихотомической зависимой переменной для тех же данных $R^2 = 0{,}318$, что согласуется с результатами, полученными для бинарной логистической регрессии.

Классификационная таблица (табл. 1.20) показывает, что в целом по построенной модели удалось правильно предсказать продолжение или непродолжение образования после среднего учебного заведения для 75,2 % выборки. Причем данная модель лучше работает для предсказания отрицательного результата, т. к. в группе тех, кто не продолжал получать образование, правильно классифицировано 83,2 % респондентов, в группе продолжавших – 63,9 %.

Таблица 1.20

Классификационная таблица для бинарной логистической регрессии

Наблюдаемые значения	Предсказанные значения		Процент правильной классификации
	нет	да	
Продолжение образования:			
нет	227	46	83,2
да	69	122	63,9
По выборке в целом	–	–	75,2

Заметим, что классификационная таблица в множественной линейной регрессии не используется, т. к. предполагается, что зависимая переменная является количественной. Однако для дихотомической зависимой переменной классификационную таблицу можно построить «вручную», сгруппировав респондентов в соответствии с предсказанной для них вероятностью наступлениия события: если вероятность наступления события меньше, чем 0,5, для респондента прогнозируется ненаступление события; если вероятность события 0,5 или больше, прогнозируется наступление события. После чего строится таблица сопряженности исходной зависимой переменной и переменной, полученной в результате группировки предсказанных значений.

Пример 1.3 (продолжение)

Воспользовавшись описанным подходом, построим таблицу классификации (табл. 1.21) для определения эффективности модели множественной линейной регрессии с дихотомической зависимой переменной (см. табл. 1.12).

Таблица 1.21

Классификационная таблица для модели множественной линейной регрессии с дихотомической зависимой переменной

Наблюдаемые значения	Предсказанные значения		Процент правильной классификации
	нет	да	
Продолжение образования:			
нет	235	38	86,1
да	76	115	60,2
По выборке в целом	–	–	75,4

Общий процент правильной классификации в двух моделях одинаков, однако «симметричность» правильных предсказаний лучше при использовании логистической модели: в группе, продолжившей получение образования, процент правильной классификации выше.

Наш методологический эксперимент показал, что метод бинарной логистической регрессии статистически лучше обоснован, позволяет оце-

нить вклад категориальных независимых переменных в объясняющую способность модели в целом и определить процент правильной классификации объектов с помощью модели.

Множественная линейная регрессия с дихотомической зависимой переменной – более грубый метод исследования причинных связей, однако он намного проще для интерпретации. Таблицу правильной классификации в этом случае приходится строить дополнительно.

Мультиноминальная и порядковая логистическая регрессия. Если категориальная зависимая переменная имеет более двух градаций, они могут быть неупорядочены (номинальная переменная) или упорядочены (порядковая переменная). В первом случае применяется мультиноминальная логистическая регрессия, во втором – порядковая логистическая регрессия.

Мультиноминальная логистическая регрессия[1] следует логике *dummy*-кодирования: одно из значений зависимой переменной объявляется референтным, и для каждого из оставшихся значений строится логистическое уравнение. Таким образом, если зависимая переменная имеет k градаций, модель будет состоять из $k - 1$ уравнений. Пример мультиноминальной логистической модели можно найти в работе С. В. Сивухи и М. Х. Титмы «Социальные детерминанты самооценки успеха» (см. прил. на с. 218).

Порядковая модель логистической регрессии[2] также состоит из $k - 1$ уравнений. Пронумеруем значения зависимой переменной в порядке возрастания от 1 до k. Первое уравнение строится для минимального (первого) значения, второе – для следующего за ним и т. д., за исключением максимального значения. Каждое из уравнений предназначено для вычисления вероятности того, что зависимая переменная превысит соответствующее значение:

$$P(y > j) = \frac{e^{(B_0 + B_1x_1 + B_2x_2 + \ldots + B_kx_k)}}{1 + e^{(B_0 + B_1x_1 + B_2x_2 + \ldots + B_kx_k)}} \qquad (1.16)$$

для $j = \overline{1, k-1}$.

Очевидно, что к максимальному значению k эта формула неприложима, т. к. зависимая переменная y не может превысить это значение.

1.4. ПУТЕВОЙ АНАЛИЗ

Путевой анализ является наиболее простым вариантом подхода, получившего название моделирования структурными уравнениями. Модель путевого анализа можно рассматривать как расширение модели множественной линейной регрессии за счет добавления в нее корреляционных

[1] *Бююль А., Цёфель П.* Указ. соч. ; *Терещенко О. В., Титма М.* Дифференциация доходов в когорте тридцатилетних // Социол. журн. 1996. № 3/4.

[2] *Бююль А., Цёфель П.* Указ. соч.

и причинных связей между независимыми переменными. Таким образом, некоторые переменные обретают в путевой модели двойной статус: по отношению к одним переменным они являются зависимыми, по отношению к другим – независимыми. Переменные, рассматриваемые в модели исключительно в качестве независимых, называются экзогенными (порожденными вне системы); переменные, рассматриваемые как зависимые или имеющие двойной статус, – эндогенными (порожденными, хотя бы отчасти, внутри системы).

Пример 1.2 (продолжение)

На рис. 1.6, *а* представлена «плоская» модель множественной линейной регрессии, в которой структура корреляций между независимыми переменными не рассматривается; рис. 1.6, *б* – путевая модель, которая позволяет рассмотреть структуру связей между независимыми переменными, например, изучить опосредующее влияние образования, стажа и должности на взаимосвязь пола и зарплаты.

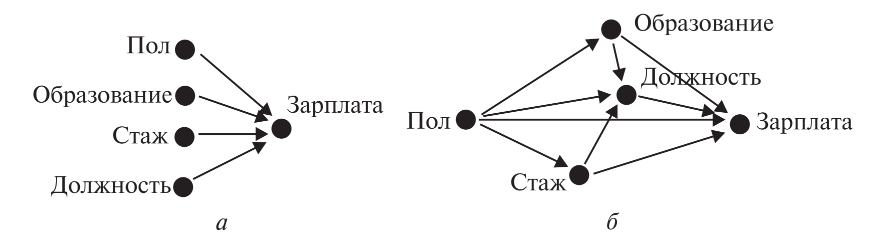


Рис. 1.6. Модель множественной линейной регрессии (*а*) и путевая модель (*б*)

Переменные «стаж», «образование» и «должность» имеют в путевой модели двойной статус. Например, «должность» одновременно выступает в качестве независимой переменной по отношению к «зарплате» и в качестве зависимой переменной по отношению к «полу», «образованию» и «стажу». «Зарплата» является только зависимой, «пол» – только независимой переменной. Таким образом, «пол» является в данной модели единственной экзогенной переменной, все остальные переменные – эндогенными.

Путевой граф является основным инструментом представления как гипотез путевого анализа, так и его результатов. Гипотезы путевого анализа касаются наличия причинных связей между переменными и их направленности (являются они прямыми или обратными). Для представления гипотез используются структурные составляющие графа: ребра (линии) – для обозначения корреляционных связей, дуги (стрелки) – для причинных связей. На граф наносятся только те связи, которые соответствуют выдвинутым гипотезам. Так, на рис. 1.6, *б* отсутствует линия или стрелка между стажем работы и образованием, поскольку наличие такой связи не предполагается.

В соответствии с гипотезами о направленности связей на соответствующих стрелках указываются знаки: «+» для прямой связи или «–» для обратной (рис. 1.7). В модель включаются также *остаточные,* или *ненаблюдаемые,* факторы – по одному на каждую эндогенную переменную, – объединяющие все не учтенные в данной модели причины, влияющие на нее. Такие факторы принято обозначать буквой *U* (от англ. *unobserved* – ненаблюдаемый).

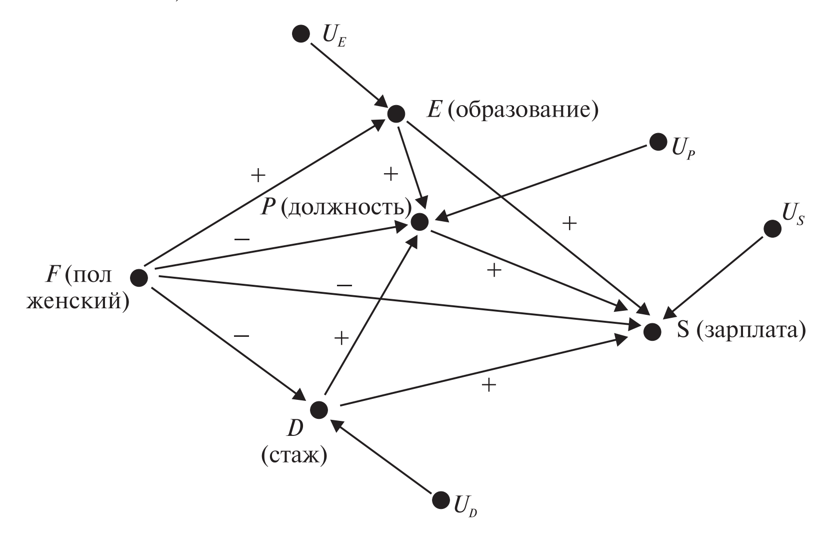


Рис. 1.7. Представление гипотез о наличии и направленности связей на графе путевого анализа

Структурные уравнения. Уравнения, описывающие структуру связей между переменными, входящими в путевую модель, называются *структурными* и позволяют проверить гипотезы, представленные посредством путевого графа. Количество уравнений равно числу эндогенных переменных – по одному на каждую. В «своем» уравнении эндогенная переменная является зависимой; в качестве независимых используются все переменные, с которыми она связана входящими стрелками. Дополнительно в уравнение включается соответствующий ненаблюдаемый фактор.

Поскольку нас интересует степень влияния переменных друг на друга, структурные уравнения представляются в стандартизированном виде. Коэффициенты, связывающие переменные в путевой модели, называются *путевыми коэффициентами*, их принято обозначать p_{YX}, где Y – зависимая переменная, X – независимая переменная. Путевые коэффициенты p_{YX} для независимых переменных X_i ($i=\overline{1,k}$) вычисляются аналогично

тому, как вычисляются стандартизированные коэффициенты в уравнении множественной линейной регрессии.

Если в структуре переменных присутствуют чисто корреляционные связи (без причинной компоненты), для них не вычисляются путевые коэффициенты, а указывается коэффициент парной корреляции.

Путевой коэффициент p_{YU_Y} для ненаблюдаемого фактора зависимой переменной U_Y вычисляется по формуле

$$p_{YU_Y} = \sqrt{1 - R^2_{Y.X_1X_2...X_k}}\ , \tag{1.17}$$

где $R_{Y.X_1X_2...X_k}$ – коэффициент множественной корреляции для зависимой переменной Y и набора независимых переменных X_1, X_2 ... X_k.

Путевые коэффициенты для ненаблюдаемых факторов, возведенные в квадрат, являются показателем неадекватности модели, они интерпретируются как доля дисперсии соответствующей зависимой переменной, *не* объясненной переменными, включенными в модель.

Пример 1.2 (продолжение)

Направленность связей между переменными показана на рис. 1.7. Все связи, представленные на рисунке, можно описать четырьмя уравнениями – по одному на каждую эндогенную переменную. Для краткости изложения обозначим переменные первыми буквами английских названий: F (female) – женский пол; D (durabilty) – стаж работы до поступления в фирму; E (education) – образование; P (position) – должность; S (salary) – начальная зарплата. Остаточные (ненаблюдаемые) факторы обозначим соответственно U_D, U_E, U_P и U_S. Пол является дихотомической переменной (1 – женский, 0 – мужской). Должность также дихотомизирована (менеджер – 1, все остальные должности – 0). Образование измеряется в годах, стаж работы – в месяцах.

Наша путевая модель состоит из 4 уравнений:

$$D = p_{DF}F + p_{DU_D}U_D\ ;$$
$$E = p_{EF}F + p_{EU_E}U_E\ ;$$
$$P = p_{PF}F + p_{PD}D + p_{PE}E + p_{PU_P}Up\ ;$$
$$S = p_{SF}F + p_{SD}D + p_{SE}E + p_{SP}P + p_{SU_S}U_s\ .$$

Для вычисления путевых коэффициентов исключим из этих уравнений ненаблюдаемые факторы:

$$D = p_{DF}F;$$
$$E = p_{EF}F;$$
$$P = p_{PF}F + p_{PD}D + p_{PE}E;$$
$$S = p_{SF}F + p_{SD}D + p_{SE}E + p_{SP}P$$

и используем тот же подход, что и при вычислении стандартизированных коэффициентов линейной регрессии (см. формулу 1.8).

Путевые коэффициенты для остаточных факторов вычисляются таким образом, чтобы сумма квадрата соответствующего путевого коэффициента с квадратом коэффициента множественной корреляции для соответствующего стандартизированного уравнения была равна 1:

$$p_{DU_D} = \sqrt{1 - r_{DF}^2};$$

$$p_{EU_E} = \sqrt{1 - r_{EF}^2};$$

$$p_{PU_P} = \sqrt{1 - R_{P.FDE}^2};$$

$$p_{SU_S} = \sqrt{1 - R_{S.FDEP}^2}.$$

Заметим, что стандартизированные уравнения для переменных D и E включают только одну независимую переменную, поэтому вместо коэффициента множественной корреляции используется коэффициент парной корреляции.

Результаты путевого анализа. Окончательная модель (результаты) путевого анализа представляется в виде графа с нанесенными на него вычисленными путевыми коэффициентами (рис. 1.8). Здесь также можно видеть путевые коэффициенты для ненаблюдаемых факторов, характеризующие качество модели.

Пример 1.2 (продолжение)

Для результирующей переменной S (зарплата) путевой коэффициент $p_{SU_S} = 0{,}55$. Будучи возведенным в квадрат, он становится показателем неадекватности модели: $0{,}55^2 = 0{,}3$, т. е. 30 % дисперсии переменной «зарплата» не объясне-

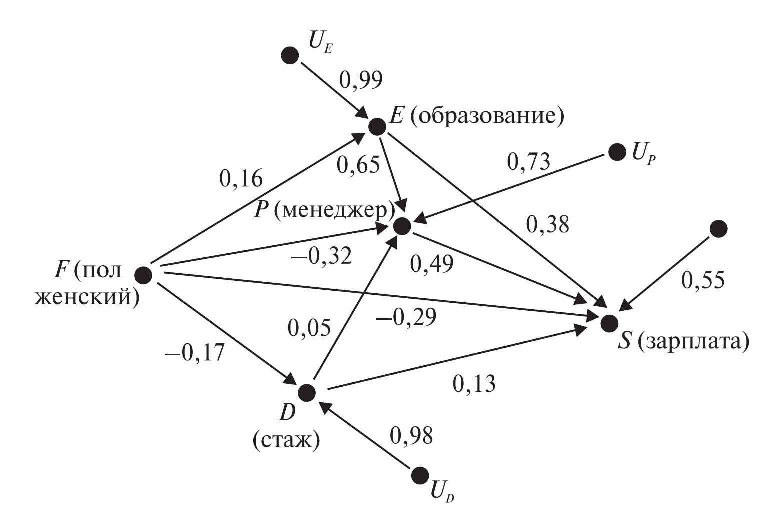


Рис. 1.8. Путевой граф с вычисленными путевыми коэффициентами

но переменными, включенными в модель, и связями между ними. Из этого следует, что модель достаточно хороша, т. к. в ее рамках объяснено 70 % дисперсии зарплаты.

Декомпозиция коэффициента парной корреляции. Помимо представления структуры связей между переменными, путевой граф может также использоваться для декомпозиции коэффициента корреляции между двумя переменными, т. е. для «разложения» его на составляющие, определения в его составе прямых и опосредованных эффектов, а также обнаружения ложных корреляций. Декомпозиция коэффициента корреляции осуществляется методом *трассирования путей*, заключающегося в том, чтобы рассмотреть все «пути» взаимодействий между переменными, вносящих свой вклад в значение рассматриваемого коэффициента.

Допустим, нас интересует коэффициент корреляции между переменными с номерами I и J. *Прямой эффект* независимой переменной I на зависимую переменную J измеряется связывающим их путевым коэффициентом p_{JI} (рис. 1.9).



Рис. 1.9. Прямой эффект независимой переменной I на зависимую переменную J

Косвенные эффекты бывают двух видов: опосредующие эффекты и эффекты ложной корреляции. *Опосредующий эффект* возникает, когда воздействие независимой переменной I на зависимую переменную J опосредуется одной или несколькими переменными, например переменной K. Значение косвенного эффекта вычисляется как произведение путевых коэффициентов по «пути» от I до J через K: $p_{JK}p_{KI}$ (рис. 1.10). Именно этот прием называется *трассированием*.



Рис. 1.10. Косвенный эффект независимой пременной I на зависимую переменную J, опосредованный переменной K

Эффект ложной корреляции возникает в случае, когда взаимодействие между переменными I и J частично или полностью обусловлено влиянием третьей переменной K (рис. 1.11). Величина ложной корреляции вычисляется как произведение путевых коэффициентов $p_{JK}p_{KI}$. Отличие ложной корреляции от опосредующих эффектов заключается в направленности причинных связей: стрелки, проходящие через вершину K, направлены

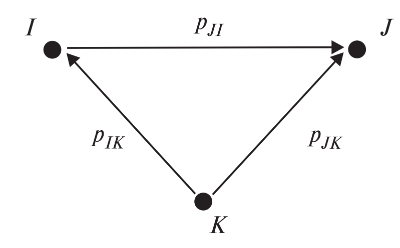


Рис. 1.11. Эффект ложной корреляции между переменными *I* и *J* под воздействием переменной *K*

в разные стороны. Тем не менее вычисление величины эффекта осуществляется так же.

Модель адекватно представляет взаимодействие между независимой переменной *I* и зависимой переменной *J*, если сумма прямого и *всех* косвенных эффектов для этих двух переменных равна коэффициенту корреляции между ними r_{JI}. Если сумма эффектов существенно отличается от значения коэффициента корреляции, модель не может рассматриваться как удовлетворительная.

Существуют определенные правила для образования легитимных «путей» между независимой и зависимой переменными при трассировании[1]. Рассмотрим их более подробно.

1. Пути «проходятся» в обратном порядке – от зависимой переменной *J* к независимой переменной *I*, т. е. от «головы» стрелки к «хвосту».

2. Если путь содержит *одну* промежуточную переменную *K*, стрелки могут менять направление, т. е. может иметь место ситуация как промежуточного эффекта, так и ложной корреляции. Путь может также включать двойную стрелку (для двух независимых переменных, коррелирующих друг с другом).

3. Если путь содержит более одной промежуточной переменной, на его легитимность существует ограничение, заключающееся в том, что «разворот» допускается только один раз. То есть если движение по стрелкам «назад» (от «головы» к «хвосту» стрелки) один раз изменилось на движение «вперед» (от «хвоста» стрелки к «голове»), повторный «разворот» уже невозможен. Что касается двойных стрелок, путь может содержать только одну такую стрелку.

Пример 1.2 (продолжение)

Рассмотрим влияние пола на начальную зарплату сотрудника в фирме (см. рис. 1.8). Полное значение коэффициента корреляции между женским полом и начальной зарплатой составляет $r_{SF} = -0{,}36$. Для декомпозиции данного коэффициента корреляции используем табл. 1.22.

В данном случае граф включает все возможные связи между переменными, поэтому неудивительно, что сумма эффектов совпала с коэффициентом корреляции.

Табл. 1.22 позволяет осуществить анализ структуры дискриминации женщин в изучаемой организации при приеме на работу, проявляющейся в назначении более низкой зарплаты, чем мужчинам.

[1] *Bohrnstedt G., Knoke D.* Statistics for Social Data Analysis. Itaca, 1988. P. 449.

Таблица 1.22

Декомпозиция коэффициента парной корреляции на прямые и косвенные эффекты

Источник корреляции	Вычисление	Значение
Прямой эффект	p_{SF}	-0,29
Косвенные эффекты: через должность через образование через образование и должность через стаж через стаж и должность	 $p_{SP}p_{PF} = 0{,}49 \times (-0{,}32)$ $p_{SE}p_{EF} = 0{,}38 \times 0{,}16$ $p_{SP}p_{PE}p_{EF} = 0{,}49 \times 0{,}65 \times 0{,}16$ $p_{SD}p_{DF} = 0{,}13 \times (-0{,}17)$ $p_{SP}p_{PD}p_{DF} = 0{,}49 \times 0{,}05 \times (-0{,}17)$	 -0,16 +0,06 +0,05 -0,02 -0,00
Сумма эффектов		-0,36
Коэффициент корреляции	r_{SF}	-0,36
Разность между коэффициентом корреляции и суммой эффектов		0,00

1) Коэффициент корреляции (–0,36) характеризует общие различия в начальной зарплате женщин и мужчин.

2) Прямой эффект пола на зарплату (–0,29) составляет большую часть коэффициента корреляции; он характеризует дискриминацию женщин непосредственно при назначении начальной зарплаты.

3) Косвенный эффект пола на зарплату через должность (–0,16) характеризует дискриминацию женщин при приеме на должность менеджера, что также сказывается на различиях в зарплате.

4) Два косвенных эффекта пола на зарплату через образование (в сумме +0,11) показывают, что женщины принимаются на более низкие должности и получают более низкую зарплату несмотря на более высокий уровень образования, чем у мужчин. Таким образом, более высокое образование женщин маскирует масштабы их дискриминации в данной фирме: если бы у женщин и мужчин был одинаковый образовательный уровень, дискриминация проявилась бы более отчетливо.

Самостоятельная работа

Пример 1.3 (продолжение)

Предположим, что поступление на бюджетную форму обучения определяется тремя основными факторами: образованием отца, образованием матери и успеваемостью респондента в средней школе. Соответствующий граф представлен на рис. 1.12. Связь между переменными «образование отца» и «образование матери» является корреляционной, она обозначается двусторонней стрелкой.

Используя путевой граф, осуществите декомпозицию коэффициента парной линейной корреляции между переменными «образование отца» и «продолжение респондентом образования».

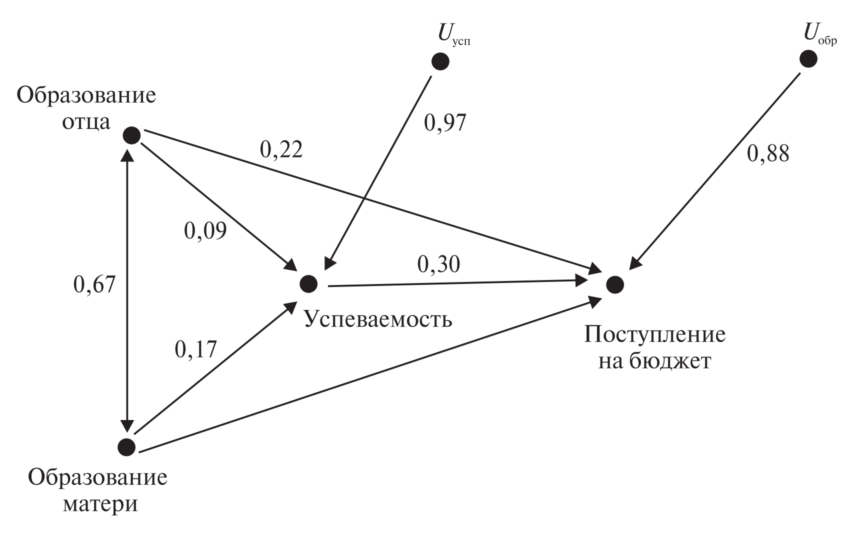


Рис. 1.12. Путевой граф с корреляционной связью между образованием отца и матери респондента

Литература

Бююль, А. SPSS: Искусство обработки информации / А. Бююль, П. Цёфель. СПб., 2001.

Гаврилец, Ю. Н. Анализ взаимосвязей, основанный на понятии структуры системы признаков / Ю. Н. Гаврилец, Г. Г. Татарова // Интерпретация и анализ данных в социологических исследованиях. М., 1987.

Елисеева, И. И. Логика прикладного статистического анализа / И. И. Елисеева, В. О. Рукавишников. М., 1982.

Елисеева И. И. Основные процедуры многомерного статистического анализа / И. И. Елисеева, Е. В. Семенова. СПб., 1993.

Елисеева, И. И. Фиктивные переменные в анализе данных / И. И. Елисеева, С. В. Курышева // Социология : 4М. 2010. № 30.

Крыштановский, А. О. Анализ социологических данных / А. О. Крыштановский. М., 2007.

Крыштановский, А. О. Ограничения метода регрессионного анализа / А. О. Крыштановский // Социология : 4М. 2000. № 12.

Куликова, А. А. Причинность в моделях латентно-структурного анализа и структурных уравнений / А. А. Куликова // Социология : 4М. 2009. № 29.

Наследов, А. Д. SPSS: Компьютерный анализ данных в психологии и социальных науках / А. Д. Наследов. СПб., 2007.

Сивуха, С. В. Социальные детерминанты самооценки успеха / С. В. Сивуха, М. Титма // Социальное расслоение возрастной когорты. Выпускники 80-х в постсоветском пространстве / отв. ред. М. Титма. М., 1997.

Терещенко, О. В. Дифференциация доходов в когорте тридцатилетних / О. В. Терещенко, М. Титма // Социол. журн. 1996. № 3/4.

Терещенко, О. В. Прикладная статистика для социальных наук: Компьютерный практикум для студентов гуманитарных специальностей / О. В. Терещенко. Минск, 2002.

Глава 2

СНИЖЕНИЕ РАЗМЕРНОСТИ

2.1. МНОГОМЕРНОЕ ПРОСТРАНСТВО ПЕРЕМЕННЫХ

Многомерное пространство переменных: геометрическая интерпретация. В количественных социологических исследованиях любой объект из изучаемой совокупности обладает множеством различных свойств, фиксируемых с помощью измеряемых переменных. Многомерность описания социологических объектов осложняет анализ данных и интерпретацию полученных результатов.

Одним из инструментов решения этой проблемы является геометрическая интерпретация набора используемых переменных, представление его в виде многомерного геометрического пространства. Переменные $x_1, x_2 \ldots x_k$ выступают в качестве осей этого пространства. Размерность пространства равна количеству переменных k. Углы между осями задаются соответствующими коэффициентами корреляций, а именно косинус угла между двумя переменными x_i и x_j ($i, j = \overline{1,k}$) равен коэффициенту корреляции между ними: $\cos(x_i, x_j) = r_{i,j}$. В частности, если коэффициент корреляции между двумя переменными равен нулю, они образуют прямой угол ($\cos 90° = 0$); коэффициент корреляции, равный +1, порождает угол, равный 0°, т. е. оси практически совпадают; коэффициент корреляции, равный –1, порождает угол, равный 180°.

Таким образом, матрица корреляций полностью описывает структуру пространства переменных: чем выше корреляция между переменными, тем ближе они расположены друг к другу (рис. 2.1). Рекомендуется рассматривать пространства из переменных, имеющих одинаковый уровень измерения – количественный, квазиинтервальный или дихотомический.

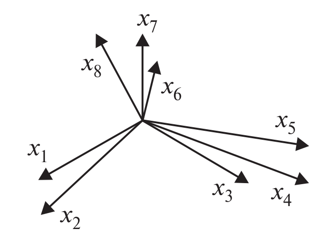


Рис. 2.1. Многомерное пространство переменных

Объекты из выборки изображаются в пространстве в виде точек, координатами которых служат значения соответствующих переменных. Рассмотрим это в наиболее простом для изображения двумерном пространстве (рис. 2.2).

Пример 1.1 (продолжение)

Построим двумерное пространство, используя в качестве осей переменные «ВВП»(валовой национальный продукт на душу населения) и «рождаемость» (на 1000 жителей). В качестве объектов в данном пространстве разместим европейские страны, для которых измерены соответствующие статистические показатели. Коэффициент корреляции между переменными $r = 0{,}08$ (см. табл. 1.4), т. е. переменные практически не коррелируют и угол между ними близок к 90°.

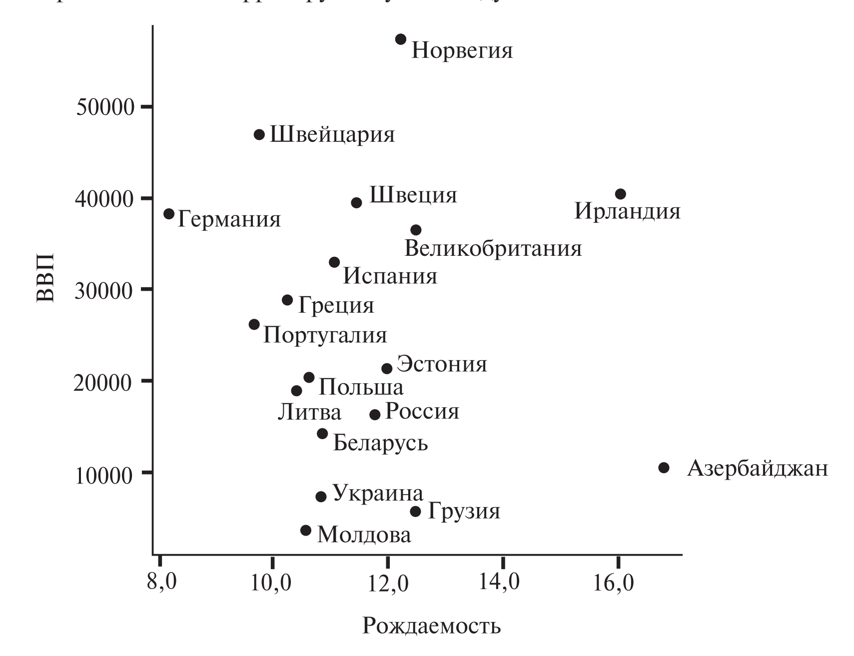


Рис. 2.2. Европейские страны[1] в пространстве двух переменных

По взаимному расположению стран в ортогональном пространстве двух переменных можно судить, например, о том, что самая высокая рождаемость в 2008 г. была в Азербайджане и Ирландии, самая низкая – в Германии, Швейцарии и Португалии. Самый высокий ВВП на душу населения был в Норвегии и Швейцарии, самый низкий – в Молдове, Грузии и Украине. В группе восточноевропейских стран Беларусь имела средний уровень ВВП на душу населения и средний уровень рождаемости.

[1] На графике представлены не все европейские страны, чтобы сделать его лучше читаемым.

Расстояния между объектами в пространстве переменных. Матрица расстояний. Между объектами в геометрическом пространстве переменных могут вычисляться расстояния. В отличие от корреляции, которая является мерой сходства, расстояние является мерой различий: чем больше расстояние между объектами, тем сильнее они отличаются друг от друга по значениям переменных, например по ответам на вопросы анкеты. Два объекта идентичны, если описывающие их переменные принимают одинаковые значения; в этом случае расстояние между ними равно нулю. Таким образом, в социальных науках расстояние между объектами может интерпретироваться как социальная дистанция.

Свойства расстояния между двумя объектами A и B:

1) $d_{A,B} \geq 0$;

2) расстояние $d_{A,B} = 0$, если объекты A и B тождественны друг другу (значения всех переменных для них совпадают);

3) $d_{A,B}$ может быть не ограничено «сверху»;

4) расстояние между объектами A и B симметрично: $d_{A,B} = d_{B,A}$;

5) для любых трех объектов A, B и C выполняется «неравенство треугольника»: $d_{A,B} \leq d_{A,C} + d_{B,C}$.

Полный набор расстояний между всеми парами объектов из выборки представляется в виде матрицы расстояний, которая имеет размерность $n \times n$, где n – объем выборки. На главной диагонали матрицы расстояний находятся нули; она симметрична относительно главной диагонали: $d_{A,B} = d_{B,A}$ (табл. 2.1).

Таблица 2.1

Матрица расстояний

Объект	A	B	C	...	N
A	0	$d_{A,B}$	$d_{A,C}$	...	$d_{A,N}$
B	$d_{B,A}$	0	$d_{B,C}$	...	$d_{B,N}$
C	$d_{C,A}$	$d_{C,B}$	0	...	$d_{C,N}$
...	...	...	...	...	...
N	$d_{N,A}$	$d_{N,B}$	$d_{N,C}$	...	0

Выбор меры расстояния зависит от конфигурации пространства и уровня измерения образующих его переменных. Конфигурация пространства определяется углами между осями. В частности, пространство является *ортогональным*, если все углы в нем прямые (переменные, образующие пространство, не коррелируют друг с другом).

К числу наиболее часто используемых мер расстояния относятся: многомерное евклидово расстояние; расстояние Махаланобиса для неортогональных пространств, образованных количественными и квазиколичественными переменными; расстояние Хемминга («city-block») для

пространства, образованного дихотомическими переменными; расстояние Чебышева для пространства, образованного порядковыми шкалами одинаковой размерности.

Многомерное расстояние Евклида является обобщением двумерного расстояния Евклида на пространства бóльшей размерности. Оно применяется для ортогональных пространств, образованных более чем двумя количественными или квазиколичественными переменными:

$$d_{A,B} = \sqrt{\sum_{i=1}^{k} (x_i(A) - x_i(B))^2}\ , \tag{2.1}$$

где $d_{A,B}$ – расстояние между объектами A и B; $x_i(A)$, $x_i(B)$ – значения переменной x_i для объектов A и B; k – количество переменных в пространстве.

Пример 1.1 (продолжение)

Пространство, изображенное на рис. 2.2, является двумерным, поэтому расстояния Евклида для расположенных в нем стран вычисляются при $k = 2$ (табл. 2.2).

В ортогональном пространстве двух переменных Беларусь наиболее близка к Литве (расстояние 1265) и наиболее удалена от Швейцарии (25093).

Вычисление расстояния, например, между Беларусью и Литвой осуществляется следующим образом:

$$d_{BY,LT} = \sqrt{\left(GNP(BY) - GNP(LT)\right)^2 + \left(BR(BY) - BR(LT)\right)^2} =$$
$$= \sqrt{(12607 - 11342)^2 + (11{,}12 - 10{,}64)^2} = \sqrt{1265^2 + 0{,}48^2} = 1265,$$

где *GNP (Gross National Product)* – валовой национальный продукт; *BR (Birth Rate)* – рождаемость; *BY* – Беларусь; *LT* – Литва. Очевидно, что основной вклад в расстояние в данном случае вносят различия в ВВП.

Если пространство количественных переменных не является ортогональным, используется ***расстояние Махаланобиса*** D^2, которое также является обобщением евклидова расстояния[1]: если коэффициенты корреляции между всеми переменными равны нулю, расстояние Махаланобиса эквивалентно квадрату евклидова расстояния.

Для дихотомических переменных вычисляется ***расстояние Хемминга*** (синонимы: «city-block», «расстояние городских кварталов», «манхэттенское» расстояние), которое равно количеству несовпадений ответов респондентов A и B по набору дихотомических переменных:

$$d_{A,B} = \sum_{i=1}^{k} \left| x_i(A) - x_i(B) \right|. \tag{2.2}$$

[1] Мы не приводим здесь формулу расстояния Махаланобиса из-за ее чрезвычайной громоздкости. См.: *Афифи А., Эйзен С.* Статистический анализ: подход с использованием ЭВМ. М., 1982. С. 330.

Таблица 2.2

Матрица расстояний Евклида, вычисленных по переменным «ВВП» и «рождаемость» для европейских стран

Страна	Австрия	Беларусь	Велико-британия	Латвия	Литва	Польша	Россия	Украина	Фин-ляндия	Швей-цария	Эстония
Австрия	0	11524	389	9315	12789	13971	15020	19128	213	13569	4180
Беларусь	11524	0	11135	2209	1265	2447	3496	7604	11737	25093	7344
Велико-британия	389	11135	0	8926	12400	13582	14631	18739	602	13958	3791
Латвия	9315	2209	8926	0	3474	4656	5705	9813	9528	22884	5135
Литва	12789	1265	12400	3474	0	1182	2231	6339	13002	26358	8609
Польша	13971	2447	13582	4656	1182	0	1049	5157	14184	27540	9791
Россия	15020	3496	14631	5705	2231	1049	0	4108	15233	28589	10840
Украина	19128	7604	18739	9813	6339	5157	4108	0	19341	32697	14948
Финляндия	213	11737	602	9528	13002	14184	15233	19341	0	13356	4393
Швейцария	13569	25093	13958	22884	26358	27540	28589	32697	13356	0	17749
Эстония	4180	7344	3791	5135	8609	9791	10840	14948	4393	17749	0

Пример 2.1. Вычисление расстояния Хемминга

Алгоритм вычисления расстояния Хемминга для двух объектов *A* и *B* представлен в табл. 2.3. Дихотомические переменные принимают значения 0 — «нет», 1— «да».

Таблица 2.3

Вычисление расстояния Хемминга

Переменный объект	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	Σ
A	1	1	1	1	1	0	0	0	0	0	
B	1	0	1	0	1	0	1	0	1	0	
$\lvert x_i(A)-x_i(B)\rvert$	0	1	0	1	0	0	1	0	1	0	4

Таким образом, для данного примера $d_{A,B}=\sum_{i=1}^{10}\lvert x_i(A)-x_i(B)\rvert=4.$

Расстояние Чебышева используется для порядковых шкал одинаковой размерности и представляет собой максимальную разность для двух объектов по всем переменным, взятую по абсолютной величине:

$$d_{A,B}=\max_{i=1}^{k}\lvert x_i(A)-x_i(B)\rvert. \qquad (2.3)$$

Пример 2.2. Вычисление расстояния Чебышева

Алгоритм вычисления расстояния Чебышева для двух объектов *A* и *B* представлен в табл. 2.4. Используются 7-балльные порядковые переменные.

Таблица 2.4

Вычисление расстояния Чебышева

Переменный объект	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
A	1	2	3	4	5	6	7	5	3	1
B	1	3	5	7	6	5	4	1	2	7
$\lvert x_i(A)-x_i(B)\rvert$	0	1	2	3	1	1	3	4	1	6

В данном случае $d_{A,B}=\max_{i=1}^{10}\lvert x_i(A)-x_i(B)\rvert=6$.

Заметим, что расстояния Хемминга и Чебышева используются как для ортогональных, так и для неортогональных пространств.

Снижение размерности пространства переменных: постановка задачи. Одновременный анализ большого числа переменных, в той или иной степени коррелирующих между собой, вызывает значительные затруднения. Задача снижения размерности заключается в том, чтобы уменьшить число анализируемых переменных, сохранив при этом большую часть ис-

ходной информации, и по возможности добиться ортогональности нового пространства, т. к. большинство методов классификации и изучения причинных связей изначально разработаны для некоррелирующих друг с другом переменных.

Все процедуры снижения размерности основаны на идее о том, что тесно коррелирующие между собой переменные измеряют близкие по содержанию свойства объектов из выборки и, соответственно, могут быть объединены (интегрированы) в более «крупные» переменные. В многомерном пространстве оси зачастую расположены неравномерно, они образуют «блоки», внутри которых переменные тесно коррелируют друг с другом. Исследование структуры пространства заключается в том, чтобы выделить эти блоки и проинтерпретировать их содержательно. Каждая из проинтерпретированных групп связанных между собой переменных может быть заменена одной новой переменной (рис. 2.3), что и приводит к снижению размерности.

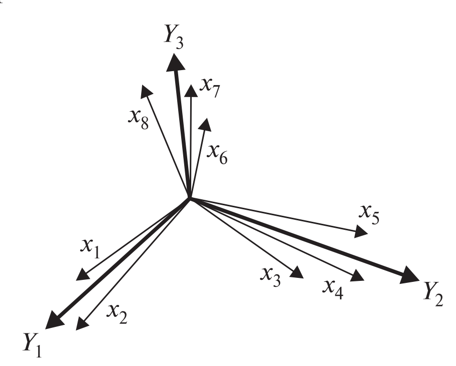


Рис. 2.3. Снижение размерности пространства переменных

Методами снижения размерности могут решаться две задачи: 1) исследование структуры переменных как самостоятельная задача при изучении ценностей, мотивов, оценок и т. п., а также при проверке внутренней валидности тестов и других конструируемых индексов; 2) собственно снижение размерности, т. е. переход к пространству с меньшим количеством переменных как предварительный этап построения классификаций или моделирования причинных связей.

Основными методами снижения размерности являются метод главных компонент, факторный анализ, метод многомерного шкалирования. Они различаются теоретическими и математическими моделями конструирования новых интегральных переменных, а также вычислительными процедурами.

Самостоятельная работа

Изобразите пространство переменных, соответствующее матрице корреляций (табл. 2.5), в которой представлены переменные, измеряющие предпочтения при выборе горнолыжного курорта. Используются 5-балльные шкалы: 1 – совершенно не важно; 5 – очень важно.

Таблица 2.5

Матрица корреляций предпочтений при выборе курорта

Переменная	x_1	x_2	x_3	x_4
x_1 – стоимость путевки	1,00	-0,95	-0,06	-0,13
x_2 – комфорт	-0,95	1,00	-0,09	-0,04
x_3 – качество снега	-0,06	-0,09	1,00	0,99
x_4 – качество склонов	-0,13	-0,4	0,99	1,00

2.2. ИЗМЕРЕНИЕ ЛАТЕНТНЫХ ПЕРЕМЕННЫХ. СЕМАНТИЧЕСКИЙ ДИФФЕРЕНЦИАЛ

Первичные и вторичные измерения. Одна из методологических проблем социологического исследования заключается в латентности многих социологических показателей, невозможности их непосредственного (прямого) измерения. Наиболее распространенным способом решения данной проблемы является конструирование индексов – вторичных переменных, позволяющих оценить степень выраженности латентных характеристик объектов.

Индексом называется производный показатель, сконструированный из исходных переменных (*индикаторов*) посредством математических и логических операций. Основными приемами построения индексов являются:

- вычисление относительных показателей (например, процент правильных ответов на вопросы теста; индикаторами в данном случае являются общее количество вопросов в тесте и число правильных ответов);
- суммирование (например, вычисление конкурсного балла при поступлении в вуз; индикаторы – экзаменационные оценки по предметам);
- вычисление среднего арифметического (например, средний балл по аттестату зрелости; индикаторы – оценки по отдельным предметам).

В наиболее общем виде вычисление индекса, основанного на суммировании индикаторов, можно представить в виде *линейной комбинации*:

$$Y = \sum_{i}^{k} a_i X_i = a_1 X_1 + a_2 X_2 + \ldots + a_k X_k, \tag{2.4}$$

где Y – конструируемый индекс; X_i ($i=\overline{1,k}$) – используемые индикаторы; a_i – весовые коэффициенты, отражающие относительную важность индикаторов, их «вклад» в значение индекса.

Можно легко показать, как из общей формулы (2.4) получаются частные случаи индексов. Так, если все весовые коэффициенты $a_i = 1$ ($i=\overline{1,k}$), индекс представляет собой арифметическую сумму индикаторов; если $a_i = 1/k$ ($i=\overline{1,k}$), то индекс – среднее арифметическое индикаторов.

Индекс может конструироваться как до начала исследования (на этапе программирования), так и непосредственно в ходе анализа данных. В первом случае процесс конструирования индекса включает четыре этапа: 1) перевод понятия в индикаторы посредством операциональных определений; 2) перевод индикаторов в переменные (выбор шкалы и единицы измерения); 3) перевод переменных в индекс (выбор техники конструирования, подбор весовых коэффициентов); 4) проверка индекса на надежность и валидность. Наиболее распространенным примером такого конструирования являются психометрические шкалы.

Во втором случае индекс конструируется в процессе анализа структуры связей между переменными-индикаторами, например, с помощью методов снижения размерности, однако это не освобождает аналитика от проверки полученного индекса (фактора, главной компоненты и т. п.) на надежность и валидность. Кроме того, индекс, построенный с помощью формальных методов анализа данных, нуждается в содержательной *интерпретации*. Предварительным условием построения индексов выступает изучение структуры связей между переменными-индикаторами.

Семантический дифференциал (СД) является, вероятно, наиболее «прозрачным» приложением идеи снижения размерности. Метод происходит из психосемантики, изучающей психологические аспекты восприятия человеком значений и смыслов разного рода объектов. Одна из основных задач психосемантики – построение так называемого семантического пространства, нахождение системы латентных факторов, в рамках которых «испытуемый» (респондент) оценивает окружающий мир – разного рода «объекты», включая субъектов, а также явления, понятия и др., которые в дальнейшем для простоты изложения будем называть *объектами*.

Метод предложен группой американских психологов во главе с Ч. Осгудом в 1957 г. для решения таких задач, как выявление факторов, которые определяют смысловую значимость объектов для респондента, и определение различий в восприятии респондентом разных объектов (отсюда название метода). Таким образом, первоначально СД был сфокусирован на личности респондента, а пространство, образованное выделенными факторами, рассматривалось как индивидуальное семантическое пространство, в которое респондент «помещает» объект в процессе его оценивания.

Однако в настоящее время применение данного метода значительно расширилось. Он используется в социологии, политологии, маркетинге и

смежных дисциплинах для изучения смыслов, которые респонденты придают самым разнообразным «объектам» – политическим партиям и их лидерам, торговым маркам, рекламным материалам и др.

Первоначально СД разрабатывался как проективная методика, в основе которой лежит явление синестезии – мышления по ассоциации, возникновения одних чувственных восприятий под воздействием других. Это явление отражается в любом языке, когда говорят, например, о горячем сердце, твердом характере и т. д. В частности, плохое ассоциируется с холодным, темным, низким; хорошее – с теплым, светлым, высоким и т. п.

Смыслы, которые респонденты приписывают объектам, не могут быть измерены непосредственно и поэтому рассматриваются как латентные факторы. Для их измерения в качестве исходных переменных (индикаторов) в СД используются пары прилагательных-антонимов, каждая из которых соответствует некоторому коннотативному континууму: «горячий – холодный», «хороший – плохой», «грязный – чистый» и т. д. Измерения производятся по 7-балльной шкале, например, паре «светлый – темный» соответствуют градации: очень светлый (1); светлый (2); не очень светлый (3); ни светлый, ни темный (4); не очень темный (5); темный (6); очень темный (7). *Надежность* измерений обеспечивается, как это принято в психометрических тестах, применением нескольких шкал оценки каждого фактора. Во избежание «автоматического» заполнения опросного листа респондентами шкалы, относящиеся к разным факторам, «перемешивают» и каждую вторую шкалу «переворачивают» (табл. 2.6).

Таблица 2.6

Фрагмент опросного листа семантического дифференциала (шкала «плохой – хороший» перевернута)

светлый	1	2	3	4	5	6	7	темный
плохой	1	2	3	4	5	6	7	хороший
чистый	1	2	3	4	5	6	7	грязный

Респондентам предъявляется один и тот же список объектов[1], каждый из которых оценивается по всему набору шкал. Таким образом, для каждого респондента получается индивидуальная матрица данных «объект – переменная», в которой содержатся его оценки всех тестируемых объектов по всем шкалам.

Используя факторный анализ для подтверждения *валидности* своей методики, Ч. Осгуд показал, что основу семантического пространства многих людей (свой метод, подобно многим психологам, он испытывал на студентах Чикагского университета) составляют три основных

[1] В частности, Ч. Осгуд предлагал своим студентам для оценки список из 15 понятий: любовь, ребенок, мой врач, я сам, моя работа, психическая болезнь, моя мама, здравый смысл, обман, мой супруг, самоконтроль, ненависть, мой отец, логика, секс. (См.: *Miller D. C.* Handbook of Research Design and Social Measurement. 5th ed. Newbury Park ; London ; New Delphi, 1991. P. 184.)

фактора, которые он назвал *оценкой* (шкалы «ценный – ничтожный», «чистый – грязный», «приятный – неприятный»), *силой* (шкалы «большой – маленький», «сильный – слабый», «глубокий – мелкий») и *активностью* (шкалы «быстрый – медленный», «активный – пассивный», «горячий – холодный»)[1]. Для каждого респондента и каждого оцениваемого им объекта вычисляется значение каждого фактора как среднее арифметическое входящих в него шкал (после восстановления шкал, которые предварительно были перевернуты). Таким образом, оцениваемые объекты (понятия) располагаются в индивидуальном семантическом пространстве, что и позволяет определять смыслы, приписываемые им данным респондентом.

В индивидуальном пространстве можно также изучать взаимное расположение объектов и степень близости их оценок друг к другу. Пространство, построенное Осгудом, является ортогональным, поэтому в качестве меры близости между объектами используется расстояние Евклида.

Пример 2.3. Семантический дифференциал

В табл. 2.7 отражены оценки 12 понятий по 9 исходным шкалам-индикаторам, поставленные респондентом (студентом БГУ), и значения для тех же понятий трех интегрированных переменных, представляющих оси персонального семантического пространства. Каждая из интегрированных переменных («оценка», «сила», «активность») получена в результате усреднения значений трех индикаторов, соответствующих каждой из них.

Трехмерное семантическое пространство респондента трудно изобразить графически, поэтому мы представим результаты на двух двумерных графиках: «оценка» × «сила» и «оценка» × «активность» (рис. 2.4 и 2.5).

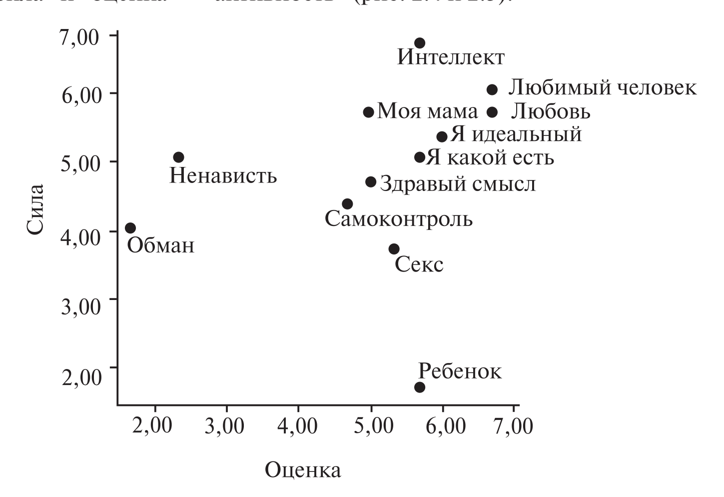


Рис. 2.4. Двенадцать понятий в семантическом пространстве респондента, образованном переменными «оценка» и «сила»

[1] *Miller D. C.* Указ. соч.

Таблица 2.7

Индивидуальная матрица данных семантического дифференциала (после восстановления «перевернутых» шкал)

Объект	Неприятный / приятный	Маленький / большой	Пассивный / активный	Грязный / чистый	Слабый / сильный	Медленный / быстрый	Ничтожный / ценный	Мелкий / глубокий	Холодный / горячий	**Оценка**	**Сила**	**Активность**
Любовь	7	6	6	6	6	4	7	5	6	6,67	5,67	5,33
Я какой есть	6	5	4	6	5	3	5	5	4	5,67	5,00	3,67
Интеллект	6	7	5	5	7	4	6	6	4	5,67	6,67	4,33
Здравый смысл	5	5	3	5	5	4	5	4	2	5,00	4,67	3,00
Самоконтроль	4	4	2	5	5	3	5	6	1	4,67	5,00	2,00
Моя мама	6	6	3	6	6	2	6	4	2	6,00	5,33	2,33
Я идеальный	6	6	4	6	6	5	6	4	4	6,00	5,33	4,33
Любимый человек	7	6	5	6	7	5	7	5	4	6,67	6,00	4,67
Секс	5	4	4	5	4	4	6	3	5	5,33	3,67	4,33
Ребенок	6	1	7	5	2	6	6	2	5	5,67	1,67	6,00
Обман	1	4	4	2	5	5	2	3	4	1,67	4,00	4,33
Ненависть	2	5	4	3	6	5	2	4	3	2,33	5,00	4,00

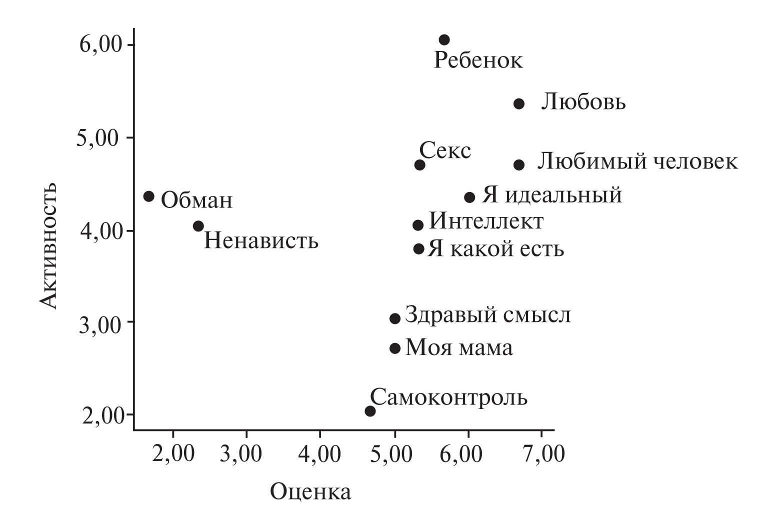


Рис. 2.5. Двенадцать понятий в семантическом пространстве респондента, образованном переменными «оценка» и «активность»

Использование семантического дифференциала в массовых опросах. При переходе от психосемантического к социологическому использованию СД возникает методологическая проблема, связанная с организацией данных, которые имеют три измерения – респондент, объект, переменная. Например, если опрошены 500 человек, каждый из которых оценил 10 объектов по 20 шкалам, то общая размерность массива данных составляет 500 × 10 × 20 = 100 000 чисел. На каждого респондента в таком массиве, в отличие от «традиционной» матрицы данных, приходится 10 записей (по числу оцененных объектов), в каждой из которых содержатся значения 20 переменных.

Задачи социологического исследования методом СД обычно связаны не с индивидуальными особенностями восприятия, а с изучением «образа» оцениваемых объектов в глазах группы респондентов. Для их решения значения факторов, полученные для каждого объекта от разных респондентов, усредняются.

В пространстве семантического дифференциала можно также «разместить» респондентов в соответствии с оценками, выставленными ими одному из объектов. В этом случае респондентов можно классифицировать по отношению к соответствующему объекту, например, электорат одного из кандидатов или лояльные потребители торговой марки.

При использовании СД в непсихологических целях методики не всегда бывают проективными. Другими словами, могут измеряться не скрытые смыслы, а конкретные характеристики оцениваемых объектов. Неизмен-

ными остаются методические приемы — 7-балльные (иногда 9-балльные) биполярные шкалы-индикаторы, вычисление интегрированных индексов как среднего арифметического входящих в них шкал, проверка их валидности методами факторного анализа.

Проверка валидности новых методик на базе СД заключается в том, что шкалы, относящиеся к одному индексу, проверяются на взаимную коррелированность. Для анализа структуры измеряемых шкал обычно используются методы снижения размерности (факторный анализ, метод главных компонент). Пример такого подхода можно найти в работе Ю. Л. Качанова и И. В. Задорина «Структура личностного образа народного депутата в сознании избирателей г. Москвы» (см. прил. на с. 187).

2.3. МЕТОД ГЛАВНЫХ КОМПОНЕНТ

Модели метода главных компонент. Метод главных компонент (МГК) — исторически первый подход к снижению размерности пространства переменных. Метод был разработан английским статистиком К. Пирсоном в 1901 г. для проверки валидности психометрических шкал. Основная идея метода базируется на геометрической интерпретации пространства переменных и состоит в том, чтобы выделить в многомерном пространстве группы тесно коррелирующих между собой переменных и заменить их интегральными индексами (главными компонентами), которые сохранили бы большую часть исходной информации. Другими словами, МГК позволяет заменить набор из k исходных переменных $x_1, x_2 \dots x_k$ набором из l новых переменных (главных компонент) $y_1, y_2 \dots y_l$, причем $l << k$, и сохранить при этом бóльшую часть исходной информации.

Сегодня в большинстве случаев МГК используется на первом этапе факторного анализа, несмотря на то, что модель главных компонент существенно отличается от модели факторного анализа. Метод главных компонент является скорее геометрическим, нежели статистическим, и реализует процесс снижения размерности через конструирование новых (интегрированных) переменных посредством линейных преобразований исходного пространства. В то время как модель факторного анализа предполагает существование скрытых (латентных) переменных, которые не могут быть измерены непосредственно, но могут быть выявлены и измерены посредством анализа структуры связей между исходными переменными.

Геометрическая модель МГК показывает, как главные компоненты конструируются из исходных переменных (рис. 2.6). Толщина линий показывает величину вклада исходных переменных в главные компоненты. Так, в главную компоненту y_1 наибольший вклад вносят переменные x_1, x_2 и x_3, в главную компоненту y_2 — переменные x_4 и x_5, в главную компоненту y_l — переменные x_{k-2}, x_{k-1}, x_k.

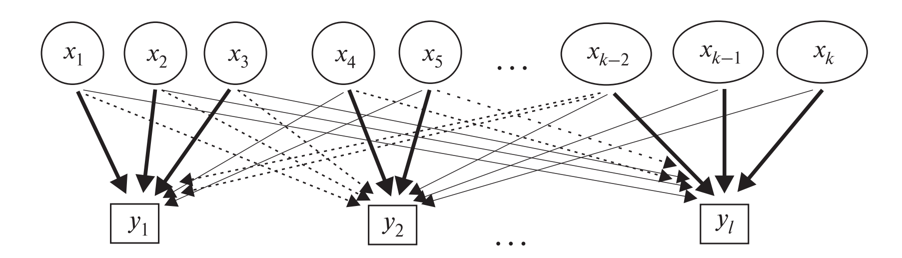


Рис. 2.6. Модель метода главных компонент:

▬ значительный вклад; — вклад средней величины;
--- незначительный вклад

Для конструирования главных компонент используется линейная модель в стандартизированных переменных:

$$y_i = \sum_{j=1}^{k} a_{i,j} z_j, \tag{2.5}$$

где y_i – главная компонента с номером $i\ (i = \overline{1,l}\ )$; z_j – стандартизированная исходная переменная x_j ($z_j = \left(x_j - \overline{x}_j\right) / s_j$, $j = \overline{1,k}$); $a_{i,j}$ – коэффициент, отражающий вклад переменной z_j в главную компоненту y_i. Заметим, что формула (2.5) фактически совпадает с формулой для вычисления индекса как линейной комбинации индикаторов (2.4); единственное различие состоит в том, что в данном случае индикаторы (исходные переменные) предварительно стандартизируются.

Линейная модель конструирования главных компонент из стандартизированных исходных переменных позволяет использовать в методе главных компонент количественные, квазиинтервальные и дихотомические переменные, т. к. для всех этих видов переменных возможно вычисление среднего арифметического и дисперсии, а следовательно, и стандартизация.

Проверка целесообразности применения методов главных компонент и факторного анализа. Методы снижения размерности целесообразно применять, если в структуре набора переменных существуют ярко выраженные группы. Другими словами, если корреляционные связи хотя бы между некоторыми переменными достаточно сильны. Для проверки соответствующих свойств матрицы корреляций используются тест сферичности Бартлетта и критерий адекватности выборки Кайзера – Мейера – Олкина (*KMO*).

Тест Бартлетта проверяет нулевую гипотезу о том, что рассматриваемые переменные не коррелируют друг с другом. Поэтому методы снижения размерности рекомендуется применять, если данная гипотеза может быть отвергнута, т. е. значение статистики Бартлетта достаточно велико и вероятность ошибки I рода (неправильно отвергнуть нулевую гипотезу) мала ($p \le 0{,}05$).

Значение *критерия КМО* увеличивается, приближаясь к 1, по мере возрастания целесообразности применения факторного анализа. Так, М. Норусис образно характеризует значение $KMO \geq 0{,}9$ как невероятное, $KMO \geq 0{,}8$ – заслуживающее одобрения, $KMO \geq 0{,}7$ – среднее, $KMO \geq 0{,}6$ – посредственное, $KMO \geq 0{,}5$ – мизерабельное. Значение $KMO \leq 0{,}5$ представляется неприемлемым[1].

Заметим, что МГК, являющийся скорее геометрическим, чем статистическим методом, может применяться, даже если предварительная проверка двух этих критериев привела к отрицательным результатам. Однако для других методов первоначального выделения факторов выполнение данных критериев считается обязательным.

Отметим также, что критерии Бартлетта и *КМО* не могут быть вычислены, если хотя бы одна из исходных переменных является линейной комбинацией других. В этом случае рекомендуется исключить из анализа данных одну из этих переменных.

Алгоритм МГК состоит из нескольких последовательных этапов.

Первый этап – стандартизация исходных переменных, т. е. переход от исходного пространства переменных $x_1, x_2 \ldots x_k$ к пространству стандартизированных переменных $z_1, z_2 \ldots z_k$ по формуле вычисления z-оценок: $z_j = \left(x_j - \overline{x}_j\right) \big/ s_j,\ \ j = \overline{1,k}$.

Среднее арифметическое стандартизированных переменных равно нулю ($\overline{z}_j = 0$), дисперсия и стандартное отклонение равны единице ($s_j^2 = s_j = 1$). Это приводит к двум важным результатам. Во-первых, с точки зрения геометрической интерпретации, осуществляется «перенос» начала координат в центр «облака» данных (рис. 2.7). Во-вторых, дисперсия всех переменных становится одинаковой (равной 1). Многие методы статистического анализа данных, включая МГК, используют дисперсию в качестве меры информативности. Соответственно, стандартизированные переменные $z_1, z_2 \ldots z_k$ имеют одинаковую информативность, а суммарный объем заключающейся в них информации равен k $\left(\sum_{j=1}^{k} s_j^2 = k\right)$.

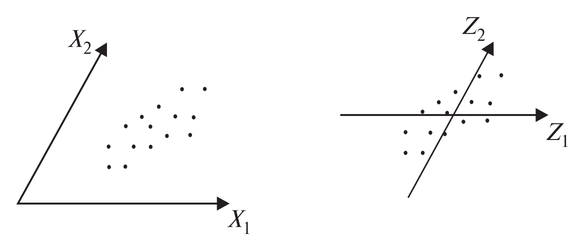


Рис. 2.7. Стандартизация переменных (перенос начала координат)

[1] *Norusis M.* SPSS Professional Statistics. Chicago, 1994. P. 50–53.

Второй этап — линейное преобразование пространства стандартизированных переменных с целью построения нового ортогонального пространства $y_1, y_2 \dots y_k$. Линейное преобразование осуществляется по формуле

$$y_i = \sum_{j=1}^{k} a_{i,j} z_j \quad (i = \overline{1,k}), \tag{2.6}$$

где y_i — новая переменная с номером i ($i = \overline{1,k}$); z_j — стандартизированная переменная с номером j ($j = \overline{1,k}$); $a_{i,j}$ — коэффициенты перехода от набора переменных $z_1, z_2 \dots z_k$ к набору переменных $y_1, y_2 \dots y_k$.

Более подробно формулу (2.6) можно представить в виде системы линейных уравнений:

$$\begin{aligned} y_1 &= a_{1,1} z_1 + a_{1,2} z_2 + \dots + a_{1,k} z_k; \\ y_2 &= a_{2,1} z_1 + a_{2,2} z_2 + \dots + a_{2,k} z_k; \\ &\dots \\ y_k &= a_{k,1} z_1 + a_{k,2} z_2 + \dots + a_{k,k} z_k. \end{aligned} \tag{2.7}$$

Коэффициенты $a_{i,j}$ вычисляются таким образом, чтобы выполнялись следующие условия[1].

1) Дисперсии новых переменных y_i численно равны собственным значениям[2] исходной матрицы корреляций $s^2(y_i) = \lambda_i$. Сумма собственных значений матрицы корреляций $\sum_{i=1}^{k} \lambda_i = k$, следовательно, информация, содержавшаяся в наборе стандартизированных переменных $z_1, z_2 \dots z_k$, полностью сохраняется в наборе новых переменных $y_1, y_2 \dots y_k$.

2) Переменные $y_1, y_2 \dots y_k$ пронумерованы в порядке убывания дисперсий: $s^2(y_1) \geq s^2(y_2) \geq \dots \geq s^2(y_k)$.

3) Переменные y_i ($i = \overline{1,k}$) ортогональны, т. е. не коррелируют друг с другом.

Таким образом, получено новое пространство переменных $y_1, y_2 \dots y_k$, размерность которого совпадает с размерностью исходного пространства. Новое пространство ортогонально, и переменные в нем упорядочены по убыванию дисперсии, т. е. по убыванию их информационной емкости: наиболее информационно ценные переменные имеют первые номера, наименее ценные — последние. Другими словами, мы «развернули» пространство стандартизированных переменных, сделав его ортогональным и одновременно упорядочив переменные по убыванию дисперсии

---

[1] Более подробно см.: *Ким Дж.-О., Мьюллер Ч. У.* Факторный анализ: Статистические методы и практические вопросы // Факторный, дискриминантный и кластерный анализ / под ред. И. С. Енюкова. М., 1989. С. 16.

[2] Характеристические числа квадратной симметричной матрицы размерности $k \times k$, вычисляемые единственным способом, упорядоченные по убыванию, в сумме дающие число k.

(рис. 2.8). Заметим, что средние арифметические новых переменных равны нулю: $\overline{y}_i = 0$ $(i = \overline{1,k})$, а их суммарная дисперсия $\sum_{i=1}^{k} s^2(y_i) = k$.

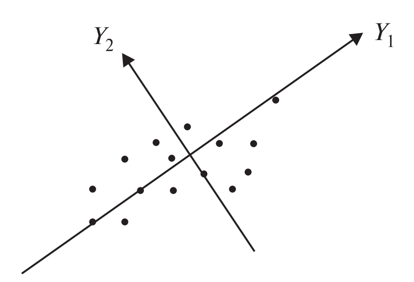


Рис. 2.8. Ортогональное пространство после «разворота» осей

Третий этап – определение числа главных компонент. Суть МГК состоит в том, чтобы сократить размерность пространства переменных посредством «отсечения» некоторого количества наименее информативных переменных с максимальными номерами. Оставшиеся l переменных $y_1, y_2 \ldots y_l$ $(l \ll k)$ называются *главными компонентами*. В процессе выделения главных компонент система уравнений (2.6) сокращается до l уравнений:

$$y_i = \sum_{j=1}^{k} a_{i,j} z_j \quad (i = \overline{1,l}), \tag{2.8}$$

Очевидно, что сумма дисперсий главных компонент меньше, чем сумма дисперсий исходных переменных (k). Следовательно, «платой» за сокращение размерности является потеря определенной части информации. Долю сохраненной информации можно определить по формуле $I = \sum_{i=1}^{l} \lambda_i \Big/ k$. Соответственно доля утраченной информации составляет $1 - I = \sum_{i=l+1}^{k} \lambda_i \Big/ k$.

Ключевым моментом выделения главных компонент является определение их количества (l). Эта задача не имеет однозначного решения. Можно использовать следующие критерии для определения числа главных компонент[1].

Критерий, основанный на собственных числах матрицы корреляции, заключается в том, чтобы ограничить отбор главных компонент теми переменными y_i, которым соответствуют собственные значения $\lambda_i \geq 1$, т. к. их информационная ценность ($s^2(y_i) \geq 1$) заведомо выше информационной ценности отсеченных переменных ($s^2(y_i) < 1$).

[1] *Ким Дж.-О., Мьюллер Ч. У.* Указ. соч. С. 35–39.

Критерий, основанный на доле сохраненной дисперсии, состоит в том, чтобы суммарная дисперсия главных компонент составляла не менее заданной доли исходной суммы дисперсий k. При применении данного критерия рекомендуется использовать накопленные относительные величины собственных значений:

$$\lambda_1/k\,;\ (\lambda_1+\lambda_2)/k;\ \ldots\ (\lambda_1+\lambda_2+\ldots+\lambda_l)/k.$$

Последнее значение должно быть не меньше заданного.

Критерий Каттелла (*графический критерий, критерий «каменной осыпи»*) использует график, в котором по оси абсцисс откладываются номера собственных значений 1, 2 ... k, а по оси ординат – сами собственные значения $\lambda_1, \lambda_2 \ldots \lambda_k$. Правая часть графика представляет прямую, а левая – нелинейную убывающую кривую. Точка пересечения этих линий, в которой график переходит в прямую линию, дает номер последней интерпретируемой компоненты.

Таким образом, все подходы к определению количества главных компонент основаны на дисперсии «повернутых» переменных y_1, y_2 ... y_k, а она, в свою очередь, определяется формальными характеристиками матрицы корреляций – ее собственными значениями $\lambda_1, \lambda_2 \ldots \lambda_k$.

Пример 1.1 (продолжение)

Для примера со статистическими показателями европейских стран собственные значения матрицы корреляций имеют следующие значения (табл. 2.8).

Таблица 2.8

Собственные значения матрицы корреляций

Компонента	Объясненная дисперсия		
	λ_i	% дисперсии	накопленный %
y_1	3,82	54,60	54,60
y_2	2,32	33,12	87,72
y_3	0,45	6,37	94,09
y_4	0,19	2,70	96,79
y_5	0,15	2,15	98,94
y_6	0,06	0,83	99,77
y_7	0,02	0,23	100,00
Сумма	7,00	100,00	

Заметим, что из набора исходных переменных удалена переменная x_4 «естественный прирост», т. к. она является линейной комбинацией (разностью) двух других переменных – «рождаемость» и «смертность», что делает невозможным вычисление критериев Бартлетта и *KMO*. Таким образом, мы использовали семь исходных переменных: x_1, x_2, x_3, x_5, x_6, x_7, x_8 (см. табл. 1.2) и получили семь новых переменных-компонент: y_1 ... y_7.

Матрица корреляций для семи переменных имеет семь собственных значений, первое из которых $\lambda_1 = 3{,}82$, второе $\lambda_2 = 2{,}32$ и т. д. Их сумма равна количеству переменных $k = 7$. Дисперсии повернутых переменных (компонент) равны соответствующим собственным значениям: $s^2(y_1) = \lambda_1 = 3{,}82$ и т. д. Суммарная дисперсия семи компонент равна 7. Поэтому первая компонента сохраняет 54,6 % суммарной дисперсии, вторая компонента – 33,12 % и т. д. Соответственно, если будет оставлена одна главная компонента, сохраненная дисперсия составит 54,6 % от суммарной исходной дисперсии, если две – 87,72 % и т. д. В методах главных компонент и факторного анализа дисперсия компонент является мерой их информативности. Таким образом, две главные компоненты сохраняют почти 88 % информации, содержавшейся в 7 исходных переменных.

Обратимся теперь к критериям выбора числа главных компонент. Согласно первому критерию, сохраняются главные компоненты, у которых дисперсия $s_i^2 \geq 1$. В данном случае таких главных компонент две.

Согласно второму критерию, количество главных компонент должно быть таким, чтобы обеспечить сохранность заданного процента исходной информации. Например, если необходимо сохранить 90 % всей информации, потребуется три главные компоненты.

Чтобы воспользоваться критерием Каттелла, необходимо построить график (рис. 2.9), по оси абсцисс которого откладываются номера компонент, по оси ординат – соответствующие им собственные значения. В данном случае «каменная осыпь» переходит в прямую линию в точке, соответствующей третьей компоненте, поэтому следует анализировать три главные компоненты.

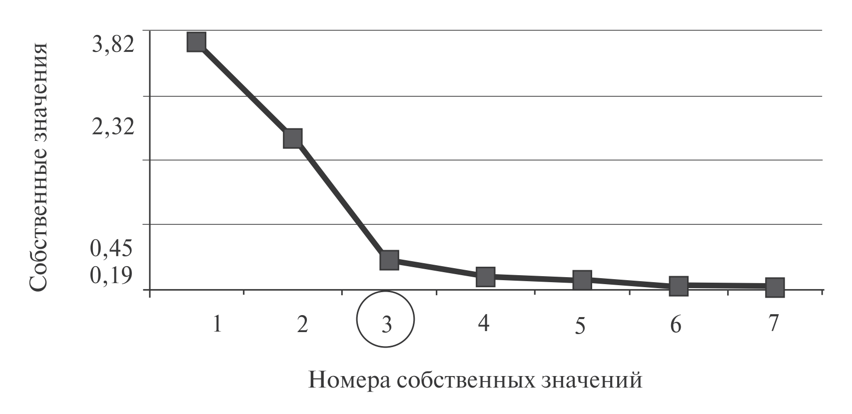


Рис. 2.9. Определение числа главных компонент по критерию Каттелла

Четвертый этап – представление результатов МГК. Результаты МГК принято представлять в виде матрицы коэффициентов линейных преобразований $a_{i,j}$ из формул (2.6) и (2.7), которые называются нагрузками на главные компоненты или факторными нагрузками (табл. 2.9). Заметим, что нагрузки нумеруются не так, как элементы большинства матриц: первый индекс здесь соответствует номеру столбца таблицы, а второй – номеру строки.

Таблица 2.9

Полная матрица нагрузок для *k* компонент

Переменная	y_1	y_2	…	y_l	…	y_k
z_1	$a_{1,1}$	$a_{2,1}$	…	$a_{l,1}$	…	$a_{k,1}$
z_2	$a_{1,2}$	$a_{2,2}$	…	$a_{l,2}$	…	$a_{k,2}$
…	…	…	…	…	…	
z_k	$a_{1,k}$	$a_{2,k}$	…	$a_{l,k}$	…	$a_{k,k}$

Полная матрица нагрузок соответствует моделям (2.6) и (2.7), полученным в результате «разворота» переменных. В табл. 2.9 и 2.10 незакрашенные l столбцов соответствуют l главным компонентам; серым цветом выделены столбцы, соответствующие отсекаемым переменным y_{l+1} … y_k.

Пример 1.1 (продолжение)

Матрица нагрузок для полного набора компонент y_1, y_2 … y_7 представлена в табл. 2.10. Для анализа оставлены первые две главные компоненты y_1 и y_2 согласно критерию, основанному на собственных значениях матрицы корреляции.

Таблица 2.10

Матрица нагрузок на компоненты

Исходная переменная	Компоненты						
	y_1	y_2	y_3	y_4	y_5	y_6	y_7
x_1 – медианный возраст	0,65	-0,72	-0,02	0,21	0,14	0,00	-0,07
x_2 – рождаемость	-0,23	0,87	0,38	0,04	0,22	0,02	-0,02
x_3 – смертность	-0,50	-0,80	0,26	0,11	0,07	0,05	0,09
x_4 – продолжительность жизни мужчин	0,91	0,30	-0,22	0,01	0,09	0,17	0,04
x_5 – продолжительность жизни женщин	0,96	0,16	-0,06	0,04	0,10	-0,16	0,06
x_6 – детская смертность	-0,76	0,49	-0,27	0,33	-0,04	-0,01	0,01
x_7 – ВВП	0,87	0,23	0,33	0,15	-0,25	0,02	0,00

Нагрузки на главные компоненты $a_{i,j}$ ($i=\overline{1,l}; j=\overline{1,k}$) выполняют в МГК три функции: 1) они используются для определения качества построенной модели; 2) служат коэффициентами корреляции между исходными переменными и главными компонентами и в этом значении используются для интерпретации последних; 3) используются как коэффициенты линейных уравнений (2.8) при вычислении значений главных компонент.

Для *определения качества модели* используются сохраненная дисперсия и общности, вычисляемые как суммы квадратов нагрузок на главные компоненты (табл. 2.11).

Таблица 2.11

Показатели качества модели: сохраненная дисперсия и общности

Стандартизированные исходные переменные	Главные компоненты				Общность
	y_1	y_2	...	y_l	
z_1	$a_{1,1}$	$a_{2,1}$	...	$a_{l,1}$	h_1^2
z_2	$a_{1,2}$	$a_{2,2}$	...	$a_{l,2}$	h_2^2
...	...	...	...	...	...
z_k	$a_{1,k}$	$a_{2,k}$	...	$a_{l,k}$	h_k^2
Дисперсия	$s_1^2=\lambda_1$	$s_2^2=\lambda_2$	...	$s_l^2=\lambda_l$	$\sum_{i=1}^{l} s_i^2=\sum_{j=1}^{k} h_j^2$
Доля сохраненной дисперсии	s_1^2/k	s_2^2/k	...	s_l^2/k	–
Накопленная доля сохраненной дисперсии	s_1^2/k	$(s_1^2+s_2^2)/k$	...	$\sum_{i=1}^{l} s_i^2 \Big/ k$	–

Сумма квадратов нагрузок по столбцу с номером i равна дисперсии главной компоненты y_i и, соответственно, собственному значению λ_i матрицы корреляций: $s_i^2=\sum_{j=1}^{k} a_{i,j}^2=\lambda_i$ $(i=\overline{1,l})$. После деления дисперсии на общий объем информации в исходном наборе переменных k получаем долю дисперсии, сохраненной в соответствующей главной компоненте s_i^2/k. В последней строке табл. 2.11 указана накопленная *доля сохраненной дисперсии* для главных компонент с соответствующими номерами.

Сумма квадратов нагрузок по строке с номером j $h_j^2=\sum_{i=1}^{l} a_{i,j}^2$ называется *общностью* исходной переменной x_j $(j=\overline{1,k})$ и показывает долю ее дисперсии, сохраненной в l главных компонентах.

Пример 1.1 (продолжение)

Оценка качества модели главных компонент для набора статистических показателей, характеризующих европейские страны, представлена в табл. 2.12.

Таблица 2.12

Сохраненная дисперсия и общности

Исходная переменная	Главные компоненты		Общность
	y_1	y_2	
x_1 – медианный возраст	0,65	-0,72	0,93
x_2 – рождаемость	-0,23	0,87	0,81
x_3 – смертность	-0,50	-0,80	0,91
x_5 – продолжительность жизни мужчин	0,91	0,30	0,92
x_6 – продолжительность жизни женщин	0,96	0,16	0,96
x_7 – детская смертность	-0,76	0,49	0,82
x_8 – ВВП	0,87	0,23	0,81
Дисперсия	3,82	2,32	6,14
% сохраненной дисперсии	54,60	33,12	
Накопленный % сохраненной дисперсии	54,60	87,72	

Дисперсия первой главной компоненты $s^2(y_1) = 3{,}82$, что составляет 54,60 % от общей дисперсии исходного набора переменных. Дисперсия второй главной компоненты $s^2(y_2) = 2{,}32$, или 33,12 % от общей дисперсии. Таким образом, общая сохраненная дисперсия составляет 87,72 %.

Общности показывают, что в максимальной степени сохранена информация из исходной переменной «ожидаемая продолжительность жизни женщин» (96 %), в минимальной степени – из переменных «рождаемость» и «ВВП» (по 81 %), что также является весьма приемлемым результатом.

Пятый этап – интерпретация главных компонент. Матрица нагрузок (см. табл. 2.9) используется также для интерпретации главных компонент. Поскольку нагрузки являются коэффициентами корреляции между главными компонентами и исходными переменными, для интерпретации используются переменные, имеющие максимальные по абсолютной величине нагрузки.

При интерпретации нагрузок различают две основные ситуации:

1) если все значительные по абсолютной величине нагрузки имеют одинаковые знаки[1], главная компонента называется *главной компонентой размера*; она показывает степень выраженности у объектов одной латентной характеристики, которую и следует интерпретировать;

2) если значительные по абсолютной величине нагрузки имеют разные знаки, то говорят о *главной компоненте формы*, которая дифференцирует

[1] Являясь коэффициентами корреляции, нагрузки могут принимать значения в интервале (–1; +1).

объекты из выборки в соответствии с наличием у них двух свойств, в некоторой степени противоположных друг другу.

С содержательной точки зрения интерпретация главных компонент происходит по ассоциации с теми исходными переменными, которые имеют на них максимальные (по абсолютной величине) нагрузки. Анализируется семантика соответствующей группы переменных, их «физический смысл». Выявляется ее общее содержание, то общее свойство, которое, по мнению исследователя, объединяет переменные в одну группу. Это свойство или группа свойств затем получает название и фигурирует в качестве названия (имени) главной компоненты[1].

Пример 1.1 (продолжение)

Для удобства интерпретации главных компонент применим распространенный прием: отсортируем исходные переменные по убыванию нагрузок (по абсолютной величине) и удалим из матрицы нагрузки, не превышающие по абсолютной величине 0,5 (табл. 2.13).

Таблица 2.13

Нагрузки, превышающие по абсолютной величине значение 0,5 и отсортированные по убыванию

Исходная переменная	Главные компоненты	
	y_1	y_2
x_6 – продолжительность жизни женщин	**0,96**	
x_5 – продолжительность жизни мужчин	**0,91**	
x_8 – ВВП	**0,87**	
x_7 – детская смертность	**-0,76**	
x_2 – рождаемость		**0,87**
x_3 – смертность		**-0,80**
x_1 – медианный возраст	0,65	**-0,72**

В нашем примере обе главные компоненты (ГК) являются компонентами формы, т. к. значительные по величине нагрузки (выделенные в табл. 2.13 жирным шрифтом) имеют разные знаки.

Первую ГК будем интерпретировать по ассоциации с такими переменными, как «продолжительность жизни мужчин» и «продолжительность жизни женщин» (средние ожидаемые), «ВВП» (со знаком «+»), «детская смертность» (со знаком «–»). На ее положительном «конце» будут располагаться страны с высоким ВВП

[1] *Викторов В. И.* Факторный анализ // Интерпретация и анализ данных в социологических исследованиях / отв. ред. В. Г. Андреенков, Ю. Н. Толстова. М., 1987. С. 223–224.

на душу населения, высокой продолжительностью жизни и низкой детской смертностью. Поскольку демографические переменные в контексте данной ГК связаны с уровнем ВВП, ее можно проинтерпретировать как «уровень благосостояния».

Вторая ГК интерпретируется в соответствии с переменными «рождаемость» (со знаком «+»), «смертность» и «медианный возраст» (со знаком «–»): на положительном ее конце будут располагаться страны с благоприятной демографической ситуацией (относительно молодое население с относительно низким медианным возрастом, низким уровнем смертности и высоким уровнем рождаемости), на отрицательном – страны с неблагоприятной демографической ситуацией (стареющее население с высоким медианным возрастом и высоким уровнем смертности, низким уровнем рождаемости). Она может быть так и названа – «демографическая ситуация».

Переменная «медианный возраст» имеет большие по величине нагрузки на обе главные компоненты, что затрудняет их интерпретацию, т. к. приходится предположить, что страны с высоким медианным возрастом имеют также высокий ВВП, и наоборот, что в целом не соответствует действительности.

Для облегчения восприятия факторной структуры может применяться графическое представление нагрузок. Для этого в пространстве двух, максимум трех, главных компонент стандартизированные исходные переменные $x_1, x_2 \ldots x_k$ изображаются в виде точек, в качестве координат используются соответствующие значения нагрузок.

Пример 1.1 (продолжение)

График нагрузок для статистических показателей, измеренных для европейских стран, представлен на рис. 2.10.

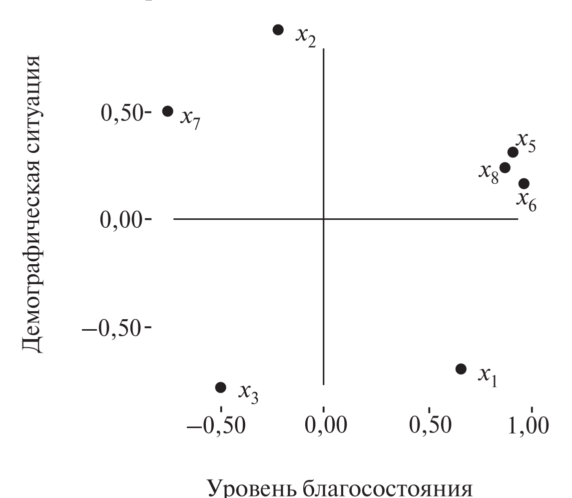


Рис. 2.10. График факторных нагрузок

Шестой этап – вычисление главных компонент. Наиболее распространенным методом вычисления главных компонент является регрессионное шкалирование, которое заключается в том, что для каждого объекта из выборки в уравнение (2.8) подставляются стандартизированные значения исходных переменных. Вычисление значений главных компонент фактически приводит к окончательному переходу из исходного пространства переменных в пространство главных компонент. Значения главных компонент, вычисленные методом регрессионного шкалирования, являются стандартизированными и обычно лежат в интервале (–4;+4).

Пример 1.1 (продолжение)

Вычисленные значения главных компонент для европейских стран представлены в табл. 2.14. Максимальные по абсолютной величине значения выделены жирным шрифтом.

Таблица 2.14

Вычисленные значения главных компонент

Страна	Медианный возраст	Рождаемость	Смертность	Продолжительность жизни мужчин	Продолжительность жизни женщин	Детская смертность	ВВП	Уровень благосостояния (y_1)	Демографическая ситуация (y_2)
Австрия	41,3	9,26	8,94	77,8	83,3	4,43	24131	1,10	-0,29
Азербайджан	28,8	17,48	6,06	71,0	76,1	51,42	8097	**-1,92**	**3,59**
Армения	32,3	12,88	8,56	70,2	76,7	31,18	11630	-1,22	1,53
Беларусь	38,3	11,12	13,80	64,5	76,2	10,24	12607	-0,99	-0,78
Бельгия	40,8	11,68	9,50	77,1	82,6	4,44	23655	0,86	0,15
Болгария	41,1	10,22	14,54	69,8	77,0	12,88	8886	-0,69	-1,20
Великобритания	39,4	12,90	9,41	77,7	81,9	5,38	23742	0,71	0,58
Венгрия	39,6	9,91	13,00	70,0	78,3	7,53	9500	-0,45	-0,97
Германия	43,7	8,32	10,30	77,6	82,7	4,21	20801	1,04	-0,99
Греция	41,4	10,47	9,56	77,7	82,4	5,45	16362	0,72	-0,23
Грузия	36,4	12,86	9,77	69,0	78,8	32,80	5984	-1,14	0,91
Дания	40,3	11,82	9,93	76,5	81,0	4,53	24621	0,68	0,13
Ирландия	33,8	16,69	6,27	77,5	82,3	5,22	27898	0,59	**2,44**

Окончание табл. 2.14

Страна	Медианный возраст	Рождаемость	Смертность	Продолжительность жизни мужчин	Продолжительность жизни женщин	Детская смертность	ВВП	Уровень благосостояния (y_1)	Демографическая ситуация (y_2)
Испания	39,5	11,33	8,57	78,0	84,3	4,33	19706	0,92	0,35
Италия	42,8	9,61	9,71	78,7	84,0	4,46	19909	1,10	-0,49
Латвия	39,8	10,39	13,48	67,0	77,8	9,66	14816	-0,56	-0,93
Литва	38,9	10,64	13,27	66,3	77,6	8,07	11342	-0,71	-0,86
Македония	35,5	11,45	9,50	72,4	76,5	16,77	4063	-0,88	0,42
Молдова	33,7	10,83	11,64	65,5	73,2	19,63	3540	**-1,72**	-0,02
Нидерданды	40,3	11,19	8,19	78,4	82,5	4,85	24695	0,99	0,33
Норвегия	38,5	12,60	8,69	78,4	83,2	3,51	28500	1,00	0,77
Польша	37,5	10,88	9,96	71,3	80,0	7,79	19160	-0,22	-0,04
Португалия	40,4	9,87	9,84	76,2	82,4	5,23	14436	0,55	-0,36
Россия	37,7	12,08	14,63	61,8	74,2	16,53	9111	-1,59	-0,66
Румыния	38,0	10,32	11,78	69,7	77,2	17,70	4895	-0,90	-0,42
Сербия	41,3	9,47	14,07	71,3	76,6	13,21	6686	-0,66	-1,30
Словакия	36,5	10,63	9,85	70,8	79,0	7,55	13033	-0,28	0,02
Словения	41,2	10,90	9,15	75,5	82,6	4,18	18170	0,71	-0,12
Украина	39,2	11,10	16,40	62,3	74,0	14,40	5003	**-1,64**	**-1,35**
Финляндия	41,8	11,23	9,26	76,5	83,3	3,31	24344	1,01	-0,03
Франция	39,7	12,76	8,53	77,9	84,9	4,10	22223	1,00	0,67
Хорватия	41,1	9,95	11,86	72,4	79,7	6,66	8904	-0,08	-0,89
Чехия	39,2	11,39	9,99	74,1	80,5	4,10	12868	0,17	-0,07
Швейцария	41,2	9,96	7,95	79,8	84,6	4,19	37700	**1,69**	0,24
Швеция	40,7	11,75	9,83	79,2	83,3	3,18	24409	1,03	0,17
Эстония	39,3	12,31	12,85	68,7	79,5	8,04	19951	-0,23	-0,29

В данном случае у нас двумерное пространство, образованное главными компонентами, поэтому мы можем изобразить его на плоскости. На рис. 2.11 представлены те же страны, что и на рис. 2.2.

Учет в структуре главных компонент, наряду с рождаемостью и ВВП, других демографических показателей (смертности, средней продолжительности жизни

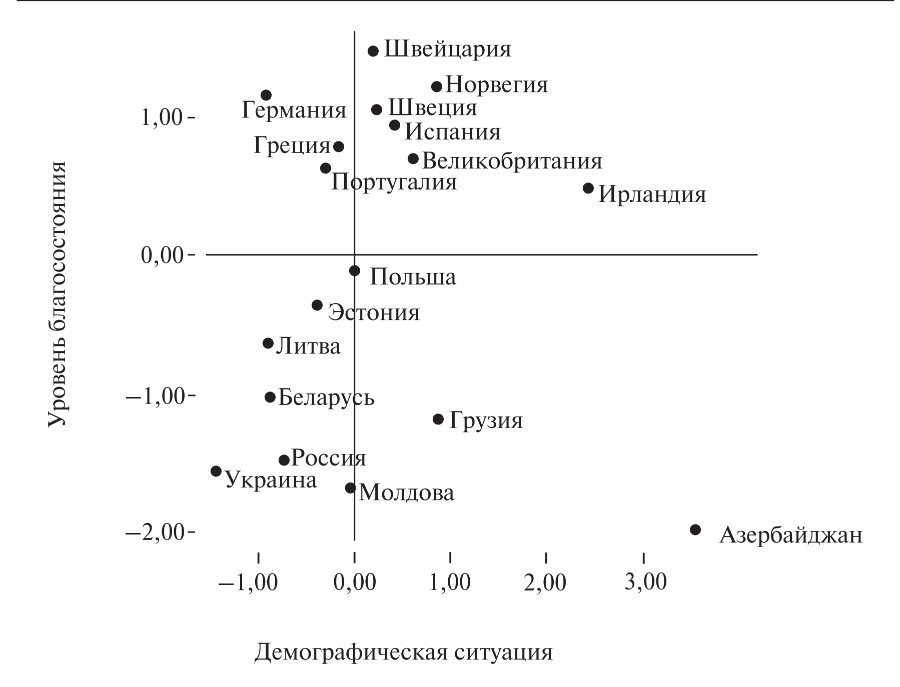


Рис. 2.11. Европейские страны в пространстве главных компонент

мужчин и женщин, медианного возраста, детской смертности) позволяет выявить ряд более общих тенденций. Так, западноевропейские страны образуют на рис. 2.11 более однородную группу, чем на рис. 2.2; их отрыв по уровню жизни от восточноевропейских стран выражен сильнее, чем по одному показателю (ВВП); восточноевропейские страны предстают более дифференцированными, особенно по уровню жизни. Интересно, например, что Россия, имевшая в 2008 г. более высокий ВВП на душу населения, чем Беларусь, в целом, с учетом таких показателей, как продолжительность жизни и детская смертность, имела более низкий уровень благосостояния, чем Беларусь.

В то же время сходный характер взаимного расположения стран на рис. 2.11 и 2.2 позволяет сделать вывод о валидности измерения и интерпретации главных компонент.

Этап вычисления значений главных компонент не является обязательным. Нередко МГК используется только для изучения структуры связей между переменными: переменные, имеющие высокие по абсолютной величине нагрузки на одну главную компоненту, могут рассматриваться как группа коррелирующих друг с другом переменных, являющаяся составной частью искомой структуры. Именно таким образом проверяется, в частности, валидность факторов семантического дифференциала.

Таковы основные этапы реализации МГК. Следует однако заметить, что главные компоненты, полученные посредством описанного выше алгоритма, не всегда удается проинтерпретировать. В этом случае исследование структуры исходных переменных можно продолжить посредством факторного анализа, в рамках которого МГК нередко применяется в качестве метода первоначального выделения факторов. В частности, результаты могут быть улучшены посредством вращения факторов, полученных методом главных компонент.

Самостоятельная работа

1. Проинтерпретируйте главные компоненты, полученные для примера с успеваемостью школьников (табл. 2.15). Используйте график нагрузок на главные компоненты.

2. Нарисуйте пространство главных компонент и представьте в нем: а) школьника, хорошо успевающего по всем предметам; б) школьника, плохо успевающего по всем предметам; в) школьника, хорошо успевающего по математическим дисциплинам и плохо – по гуманитарным; г) школьника, плохо успевающего по математическим дисциплинам и хорошо – по гуманитарным.

Таблица 2.15

Успеваемость школьников по набору дисциплин

Дисциплина (исходная переменная)	Главные компоненты	
	y_1	y_2
x_1 – русский язык	0,55	0,43
x_2 – английский язык	0,57	0,29
x_3 – история	0,39	0,45
x_4 – арифметика	0,74	-0,27
x_5 – алгебра	0,72	-0,21
x_6 – геометрия	0,60	-0,13

2.4. ФАКТОРНЫЙ АНАЛИЗ

Метод факторного анализа: общая характеристика. Факторный анализ (ФА) – один из методов снижения размерности, впервые разработанный для психологии, где представлял собой выражение идей тестового подхода.

В дальнейшем ФА стали применять не только к индивидам, но и к человеческим сообществам, а также к неодушевленным объектам для выявления их неочевидных обобщенных характеристик.

В основе ФА лежат следующие предположения[1].

1. Все наблюдаемые различия в значениях измеряемых переменных (индикаторов) обусловлены различиями в некоторых внутренних – скрытых (латентных) – свойствах испытуемых (респондентов), не поддающихся непосредственному измерению. Эти свойства получили название факторов или, точнее, *общих* факторов, влияющих на значения *всех* измеряемых переменных. Предполагается, что общие факторы непрерывны и их число относительно невелико, значительно меньше числа индикаторов, в изменении которых они проявляются. Их содержание (смысл) может быть выявлено посредством анализа структуры корреляций между измеряемыми переменными. Для отдельных респондентов можно оценить численные значения общих факторов и использовать их в дальнейшем для описания респондентов, их сравнения, классификации и т. п.

2. Каждый общий фактор имеет различную значимость для изменений разных индикаторов, т. е. существует нагрузка каждого фактора на каждую измеряемую переменную, определяющая степень такого влияния.

3. Помимо изменений, порожденных влиянием общих факторов, существуют индивидуальные изменения эмпирических переменных, вызываемые различными причинами – контролируемыми и неконтролируемыми, к которым относятся особенности инструментария, ситуация опроса, состояние респондента, личность интервьюера и др. В совокупности они составляют *хара́ктерные* факторы, по одному для каждого индикатора.

Таким образом, все изменения наблюдаемых переменных объясняются двумя группами факторов – группой изучаемых и интерпретируемых общих факторов и группой характерных факторов, которые в ФА не рассматриваются. В целом, ФА объясняет корреляции между измеряемыми переменными воздействием общих факторов.

Модели факторного анализа. Главная задача ФА – переход от некоторого числа относительно легко измеряемых характеристик изучаемого явления к относительно небольшому числу стоящих за ними латентных общих факторов. Как и при рассмотрении МГК, количество исходных переменных будем обозначать буквой k, количество общих факторов – буквой l ($l << k$). Стандартизированные исходные переменные обозначим $z_j\left(j=\overline{1,k}\right)$, общие факторы $F_i\left(i=\overline{1,l}\right)$, характерные факторы $u_j\left(j=1,k\right)$.

Структурная графическая модель факторного анализа представлена на рис. 2.12. Она иллюстрирует влияние на каждую стандартизированную исходную переменную z_j набора общих факторов F_1, F_2 ... F_l и уникального характерного фактора u_j. Корреляции между переменными z_j (и соответственно x_j) обусловливаются влиянием общих факторов (ср. с рис. 2.6).

[1] *Толстова Ю. Н.* Измерение в социологии. М., 1998. С. 90–93.

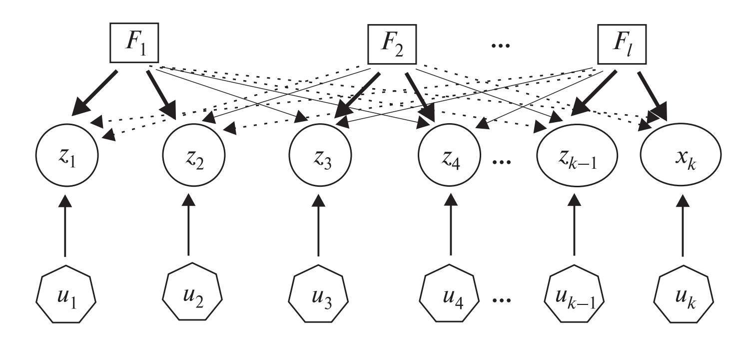


Рис. 2.12. Графическая структурная модель факторного анализа

Линейная модель факторного анализа. Графическая структурная модель ФА (см. рис. 2.12), как и в методе главных компонент, математически реализуется в виде набора линейных уравнений: каждая наблюдаемая стандартизированная переменная z_j представляется в виде линейной комбинации всех общих факторов и уникального характерного фактора:

$$z_j = \sum_{i=1}^{l} a_{i,j} F_i + u_j, \tag{2.9}$$

где z_j ($j = \overline{1,k}$) – z-оценки измеряемых индикаторов x_j; F_i ($i = \overline{1,l}$) – латентные общие факторы; $a_{i,j}$ – нагрузки факторов F_i на индикаторы z_j (факторные нагрузки); u_j ($j = \overline{1,k}$) – специфические факторы индикаторов z_j.

Количество общих факторов (l) предполагается существенно меньше количества параметров (k). В частности, Л. Терстоун предполагал, что на один фактор должны приходиться, как минимум, три наблюдаемые переменные. В целом, исследователи сходятся на том, что переменных должно быть, по меньшей мере, вдвое больше, чем факторов[1].

Линейная модель ФА позволяет понять, почему в нем не допускается использование дихотомических переменных. Действительно, никакая линейная комбинация непрерывных факторов не позволяет получить индикаторы, которые могут принимать только два значения. Таким образом, факторному анализу могут быть подвергнуты только количественные и некоторые порядковые (квазиколичественные) переменные.

Структурная модель разделения дисперсии. Дисперсия каждого из измеряемых индикаторов $s^2(x_j)$, в качестве показателя их изменчивости, состоит из двух частей – общности и характерности. *Общность* является долей дисперсии, объясненной влиянием набора общих факторов $s^2_{общ}(x_j)$, *характерность* – влиянием уникального характерного фактора $s^2_{хар}(x_j)$. В некоторых случаях характерность, в свою очередь, разделяют на *специфичность*, по-

[1] *Ким Дж.-О., Мьюллер Ч. У.* Указ. соч. С. 67.

рожденную контролируемыми источниками изменчивости индикаторов, и *остаточную дисперсию*, связанную с неконтролируемыми источниками:

$$s^2(x_j) = s^2_{общ}(x_j) + s^2_{хар}(x_j) = s^2_{общ}(x_j) + s^2_{спец}(x_j) + s^2_{ост}(x_j). \qquad (2.10)$$

Общности измеряемых переменных являются одним из показателей внутренней валидности выделенных факторов. Заметим, что терминологическое сходство модели разделения дисперсии в ФА наблюдается скорее с регрессионным анализом (см. с. 30), чем с МГК.

Алгоритм факторного анализа состоит из нескольких этапов.

Первый этап – первоначальное извлечение факторов (вычисление факторных нагрузок). Аналогично нагрузкам на главные компоненты нагрузки общих факторов являются основным результатом ФА и предметом интерпретации. Характерные факторы не интерпретируются, а рассматриваются как источники дисперсии, не объясненной влиянием общих факторов.

Для вычисления факторных нагрузок могут использоваться разные методы, в том числе (наиболее часто) МГК, учитывающий всю дисперсию исходных переменных, и группа методов общих факторов, учитывающих только общую часть дисперсии (метод главных факторов, метод максимального правдоподобия, метод наименьших квадратов, анализ образов и др.[1]). Заметим, что при выраженной структуре корреляционной матрицы (которую подтверждают высокие значения статистики Бартлетта и критерия *KMO*) результаты ФА, полученные разными методами выделения факторов, обычно очень похожи.

Общим у всех методов ФА и МГК является, во-первых, стандартизированное представление измеряемых переменных и факторов; во-вторых, представление основных результатов в виде матрицы факторных нагрузок (факторной матрицы), аналогичной матрице нагрузок на главные компоненты (см. табл. 2.11). Факторные нагрузки, как и нагрузки на главные компоненты, являются коэффициентами линейной корреляции между измеряемыми переменными и факторами и в силу этого могут использоваться для интерпретации факторов и вычисления их дисперсий, а также общностей переменных. Однако в отличие от МГК они не могут использоваться для вычисления значений факторов.

Второй этап – вращение факторной структуры. Вращение факторной структуры имеет целью максимально «упростить» факторную матрицу, чтобы облегчить ее интерпретацию. Факторная структура является *простой*, если каждая переменная-индикатор имеет только одну значительную по абсолютной величине факторную нагрузку, т. е. тесно связана только с одним фактором. Поэтому если первоначально выделенные факторы плохо поддаются интерпретации (особенно из-за того, что ма-

[1] *Ким Дж.-О., Мьюллер Ч. У.* Указ. соч. С. 18–25.

трица нагрузок не является простой), рекомендуется их дополнительное вращение.

Возможность вращения факторов основана на многозначности факторного решения, обусловленной тем, что в системе линейных уравнений (формула 2.9) количество уравнений (k) значительно больше числа неизвестных, в качестве которых выступают общие факторы (l), и система имеет бесконечное множество решений. Это позволяет выбрать оптимальное решение, удовлетворяющее таким критериям, как простота, ортогональность и интерпретируемость факторной структуры. Заметим, что ортогональность нередко является менее значимым критерием, чем интерпретируемость, и ею можно «пожертвовать», применяя неортогональные методы вращения.

Существует целый ряд методов вращения факторной структуры, ортогональных и неортогональных[1]. Ортогональные методы сохраняют при вращении некоррелированность факторов друг с другом. К ним относятся методы варимакс, квартимакс, эквимакс, биквартимакс и др. Вращение неортогональными (косоугольными) методами может привести к потере ортогональности факторов, т. е. между общими факторами могут появиться значительные по абсолютной величине значения коэффициентов корреляции. К косоугольным методам вращения относятся квартимин, коваримин, прямой облимин, облимакс, ортоблик, максплейн и др. Методы неортогонального вращения используются, когда ортогональными методами не удалось получить достаточно простую и интерпретируемую факторную структуру.

Пример 1.1 (продолжение)

Результаты вращения факторной структуры, полученной методом главных компонент (см. табл. 2.12), представлены в табл. 2.16.

Матрица нагрузок после вращения ортогональным методом варимакс стала более простой, т. к. переменная «медианный возраст», имевшая высокие нагрузки на оба фактора, теперь имеет только одну значительную по величине нагрузку.

В то же время структура факторов изменилась. В первый фактор, наряду с ожидаемой продолжительностью жизни женщин и мужчин и ВВП, теперь входит смертность (со знаком «–») и не входит детская смертность. Таким образом, уровень смертности оказался теснее связан с ВВП, чем с возрастной структурой населения, и является в большей степени социально-экономическим показателем, нежели демографическим. Детская смертность, напротив, оказалась скорее в структуре второго фактора, чем первого, и, следовательно, скорее социально-демографическим феноменом, нежели социально-экономическим. Основываясь на *знаках* факторных нагрузок, можно сказать, что в странах с высоким медианным возрастом населения и низкой рождаемостью ценность каждого рожденного ребенка выше и они, вне зависимости от уровня ВВП, изыскивают средства для борьбы с детской смертностью.

[1] *Ким Дж.-О., Мьюллер Ч. У.* Указ. соч. С. 26–35.

Таблица 2.16

Вращенная матрица факторных нагрузок

Исходная переменная	Главные компоненты	
	y_1	y_2
x_1 – медианный возраст	0,26	**0,93**
x_2 – рождаемость	0,19	**-0,88**
x_3 – смертность	**-0,82**	0,48
x_5 – продолжительность жизни мужчин	**0,95**	0,14
x_6 – продолжительность жизни женщин	**0,93**	0,29
x_7 – детская смертность	-0,46	**-0,78**
x_8 – ВВП	**0,88**	0,18

В результате вращения факторной структуры дисперсия перераспределяется между факторами более равномерно, и факторы формы нередко превращаются в факторы размера, что способствует улучшению их интерпретации.

Пример 1.1 (продолжение)

Перераспределение дисперсии между факторами после вращения представлено в табл. 2.17.

Таблица 2.17

Перераспределение дисперсии после вращения

Компонента	Собственные значения			Главные компоненты			Вращенные главные компоненты		
	дисперсия	% дисперсии	накопленный %	дисперсия	% дисперсии	накопленный %	дисперсия	% дисперсии	накопленный %
1	3,822	54,596	54,596	3,822	54,596	54,596	3,525	50,362	50,362
2	2,319	33,122	87,718	2,319	33,122	87,718	2,615	37,355	87,718
3	0,446	6,370	94,088						
4	0,189	2,703	96,791						
5	0,150	2,145	98,936						
6	0,058	0,825	99,761						
7	0,017	0,239	100,000						

Третий этап – факторное шкалирование. В случае успешной интерпретации факторов можно приступить к оценке значений факторных шкал, другими словами, к факторному шкалированию. В отличие от МГК, в факторном анализе существуют проблемы шкалирования, т. к. уравнения (2.9) не дают непосредственных оснований для вычисления значений факторов. При отсутствии «единственно правильной» методики для вычисления значений факторов используются разнообразные подходы[1]. Здесь мы рассмотрим два из них.

Регрессионное шкалирование. Если отличные от нуля нагрузки значительно различаются по величине, может использоваться регрессионное шкалирование, заключающееся в том, что для вычисления значений факторов подбираются специальные коэффициенты $b_{i,j}$ (факторные *веса*), позволяющие вычислять значения факторов как линейную комбинацию стандартизированных значений исходных переменных:

$$F_i = \sum_{j=1}^{k} b_{i,j} z_j, \tag{2.11}$$

где F_i ($i = \overline{1,l}$) – латентные факторы; z_i ($j = \overline{1,k}$) – z-оценки измеряемых индикаторов; $b_{i,j}$ – весовые коэффициенты.

В целом, этот подход аналогичен подходу, который мы рассматривали для метода главных компонент (см. с. 81). Отличие состоит в том, что весовые коэффициенты $b_{i,j}$ не являются нагрузками и не интерпретируются, а используются исключительно для вычисления значений факторов. Матрица весов имеет такую же структуру, как и матрица нагрузок. Значения факторов, вычисленные методом регрессионного шкалирования, являются стандартизированными и обычно лежат в интервале (–4; +4).

Шкалирование по неполным шкалам. Если отличные от нуля нагрузки одного фактора имеют примерно одинаковые значения, оценка значения фактора может быть получена в результате суммирования или вычисления среднего арифметического соответствующих переменных. Примером применения такого подхода является метод СД.

Понятие конфирматорного факторного анализа. Факторный анализ может быть эксплораторным (разведывательным) или конфирматорным (подтверждающим). Если гипотезы о взаимосвязях между переменными-индикаторами не включаются непосредственно в факторную модель, факторный анализ является эксплораторным. Все вышесказанное относится к этому виду ФА.

Конфирматорный ФА предназначен для проверки гипотез о структуре связей между переменными-индикаторами. Он применяется, в частности,

[1] *Ким Дж.-О., Мьюллер Ч. У.* Указ. соч. С. 52–62.

для проверки шкал психометрических тестов и подобного инструментария. Во многих случаях конфирматорный ФА предполагает специальную формализацию проверяемых гипотез в виде таблиц и / или графов.

Пример 2.3 (продолжение)

Гипотезу о структуре семантического дифференциала (СД) – три фактора, не коррелирующих между собой, по три переменные в каждом – можно формализовать в виде табл. 2.18.

Таблица 2.18

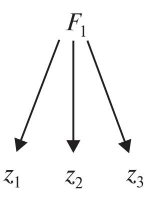
Табличное представление гипотез конфирматорного факторного анализа

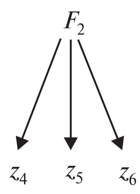
Переменная	Факторы		
	F_1	F_2	F_3
x_1	×		
x_2		×	
x_3			×
x_4	×		
x_5		×	
x_6			×
x_7	×		
x_8		×	
x_9			×

Фактор	Факторы		
	F_1	F_2	F_3
F_1	1	0	0
F_2	0	1	0
F_3	0	0	1

П р и м е ч а н и е. Слева крестиками отмечена гипотетическая коррелированность шкал СД с факторами; справа – предполагаемые корреляции между факторами.

Гипотезы конфирматорного факторного анализа могут быть также представлены в виде графа (рис. 2.13).





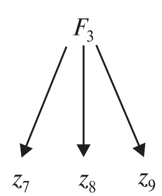


Рис. 2.13. Графическое представление гипотез конфирматорного факторного анализа

Наиболее простая гипотеза ФА – о числе общих факторов. Она не нуждается в специальной формализации и может быть проверена средствами эксплораторного ФА, в частности, она может быть подтверждена, если матрица нагрузок с заданным числом факторов является простой.

Более сложные гипотезы могут включать до трех видов информации: 1) число общих факторов; 2) природа взаимосвязи между факторами – ортогональная или косоугольная; 3) значения факторных нагрузок для каждой переменной. В некоторых случаях их проверка также может быть осуществлена посредством эксплораторного ФА. Например, можно сравнить гипотетические и наблюдаемые значения факторных нагрузок, используя статистический тест для сравнения двух коэффициентов корреляции. Однако наиболее строгая формализация и проверка гипотезы конфирматорного ФА осуществляются методами моделирования структурных уравнений и специальным программным обеспечением.

Пример 2.4 (продолжение)

Матрица повернутых факторных нагрузок (табл. 2.19) подтверждает гипотезу о структуре шкал семантического дифференциала (см. табл. 2.18, рис. 2.13).

Таблица 2.19

Матрица невращенных и вращенных факторных нагрузок

Переменная	Факторы			Вращенные факторы		
	1	2	3	1	2	3
Темный/светлый	0,71	0,66	-0,18	**0,96**	0,15	0,17
Маленький/большой	0,81	-0,37	0,44	0,15	**0,96**	-0,17
Пассивный/активный	-0,20	0,90	0,31	0,29	-0,25	**0,90**
Грязный/чистый	0,76	0,52	-0,36	**0,97**	0,12	-0,06
Слабый/сильный	0,67	-0,36	0,63	-0,02	**0,99**	-0,01
Медленный/быстрый	-0,53	0,49	0,45	-0,24	-0,25	**0,78**
Плохой/хороший	0,65	0,68	-0,24	**0,95**	0,07	0,17
Мягкий/твердый	0,86	-0,39	0,08	0,32	**0,78**	-0,44
Холодный/горячий	0,02	0,78	0,48	0,29	0,05	**0,87**

Проблемы применения факторного анализа в социологии. Ю. Н. Толстова указывает на следующие причины, ограничивающие эффективность применения ФА в социологии[1].

Во-первых, ФА разработан для количественных данных, которые достаточно редки в социологических исследованиях. Однако его часто применяют к порядковым переменным, которые можно отнести к квазиинтервальным. Дихотомические переменные могут использоваться, если первоначальное извлечение осуществляется методом главных компонент.

Во-вторых, социолог, в отличие от психолога, зачастую не имеет заранее, на этапе формирования анкеты, гипотез о латентных факторах, стоящих за наблюдаемыми переменными и объясняющих структуру связей

[1] *Толстова Ю. Н.* Указ. соч. С. 100–104.

между ними. В таких случаях уровень сохраненной факторами дисперсии и факторные нагрузки могут оказаться низкими, а факторы плохо поддаваться интерпретации. Поэтому интерпретацию результатов ФА иногда имеет смысл расценивать не как финальный этап исследования, а как этап выдвижения гипотез.

В-третьих, интерпретация результатов ФА бывает затруднена их принципиальной неоднозначностью. При той постановке задачи, которая послужила основой для разработки аппарата ФА, факторы в принципе не могут быть определены однозначно. Множество одинаково «хороших» факторных моделей может быть получено путем ротации некоторого первичного решения. Подчеркнем, что это отнюдь не должно расцениваться как недостаток метода. Напротив, в этом состоит достоинство ФА: постановка задачи была обусловлена жизненной ситуацией. Здесь мы просто сталкиваемся с принципиальной невозможностью однозначного описания социальных явлений формальными методами.

Вместе с тем, несмотря на все сказанное, тестовая традиция в социологии работает. И в настоящее время ФА успешно используется социологами, политологами, маркетологами и другими специалистами.

Самостоятельная работа

Оцените по шкалам семантического дифференциала набор понятий из примера 2.3 и проверьте гипотезу о соответствии структуры своего личного семантического пространства структуре семантического пространства, предложенного Ч. Осгудом.

2.5. МНОГОМЕРНОЕ ШКАЛИРОВАНИЕ И АНАЛИЗ СООТВЕТСТВИЙ

Многомерное шкалирование: общая характеристика. Многомерное шкалирование (МШ) — одно из современных направлений анализа данных, в рамках которого восприятия и предпочтения респондентов визуализируются в пространстве небольшой размерности. Методы МШ предназначены для решения двух основных задач: 1) снижение размерности, выявление латентных факторов, определяющих восприятия и предпочтения респондентов, лежащие в основе различий между объектами; 2) классификация объектов на основе оценки их латентных свойств.

Как и семантический дифференциал, МШ является одним из направлений математической психологии. Термин «многомерное шкалирование» был введен У. Торгерсоном в 1952 г. и в дальнейшем стал использоваться другими авторами. Бурное развитие методов МШ не в последнюю очередь связано с интенсивным совершенствованием вычислительной техники и программного обеспечения.

Исходные данные для МШ представляются в виде матрицы сходств или различий между объектами. Объекты понимаются очень широко: это могут быть оцениваемые «стимулы» (предметы, понятия, высказывания и т. п. – как в СД), физические объекты (индивиды, организации, страны, экосистемы и др.), субъекты, принимающие те или иные решения (эксперты, депутаты парламента и др.). Сходство / различия между объектами могут оцениваться респондентами непосредственно (например, методом парных сравнений), вычисляться на основе их оценки респондентами по набору шкал или некоторых объективно измеряемых переменных (статистические данные, результаты голосования в парламенте и т. п.). Таким образом, от других методов, опирающихся на представления о пространстве переменных (например, ФА), МШ отличается тем, что измеряемые отношения между объектами – точками пространства – описываются некоторыми функциями сходства / различий, которые заданы для всех пар точек пространства и не обязательно являются результатом вычислений.

В дальнейшем при обозначении сходства / различий мы будем для простоты использовать термин *расстояние*, во-первых, по аналогии с мерой различий, применяемой в геометрическом пространстве, во-вторых, потому, что мера различия всегда может быть получена из меры сходства как обратная к ней функция.

Основным инструментом МШ является построение координатного пространства небольшой размерности (одно, два, максимум три измерения) и визуализация в нем структуры расстояний между изучаемыми объектами. Другими словами, это построение *карты*[1], которая в зависимости от природы данных может являться картой восприятий, предпочтений или некоторых объективных различий между объектами. При построении карты подбирается конфигурация точек, расстояния между которыми в максимальной степени соответствуют исходной матрице расстояний[2]. Эта простая геометрическая модель приводит к содержательно интерпретируемому решению: оси построенного пространства несут определенную смысловую нагрузку, они рассматриваются как латентные факторы – *шкалы*, лежащие в основе объективного или субъективного сходства объектов. Каждый объект в той или иной степени характеризуется значениями этих шкал, и таким образом существует возможность оценить свойства отдельных объектов и получить представление об их структуре.

Способность МШ визуализировать структуру расстояний между объектами легко проиллюстрировать следующим примером.

---

[1] Впоследствии картами стали называться любые изображения объектов в пространстве переменных, полученные не только методами МШ.

[2] Очевидно, что точного соответствия добиться практически невозможно, поэтому речь идет о максимально достижимой степени соответствия.

Пример 2.4. Карта Беларуси

В табл. 2.20 представлена матрица расстояний между областными центрами Республики Беларусь по карте (в форме нижнего треугольника).

Таблица 2.20

Матрица расстояний между областными центрами Беларуси (км)

Объект	Брест	Витебск	Гомель	Гродно	Минск	Могилев
Брест	0					
Витебск	555	0				
Гомель	485	310	0			
Гродно	204	458	492	0		
Минск	330	228	282	258	0	
Могилев	490	145	168	434	180	0

На рис. 2.14 эта матрица визуализирована средствами МШ. Легко убедиться в высокой степени соответствия графической структуры географической карте страны.

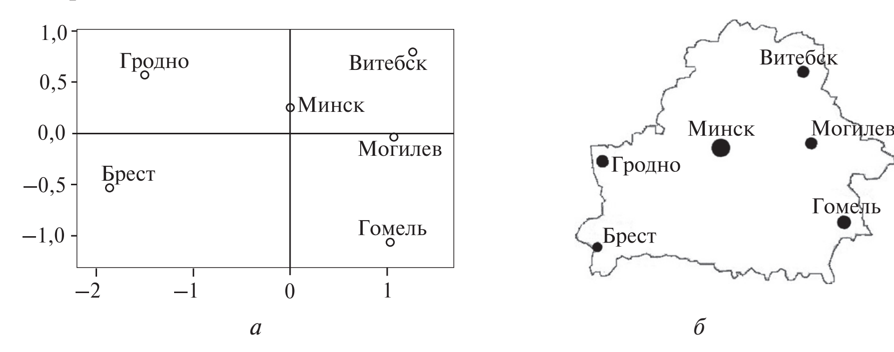


Рис. 2.14. Визуализация матрицы расстояний методом МШ между областными центрами Беларуси (*а*) и географическая карта (*б*)

Общими чертами методов МШ являются визуальное представление результатов, универсальность моделей и вычислительных процедур, возможность использования разнотипной информации.

В зависимости от используемых измерительных шкал применяются разнообразные процедуры определения расстояний между объектами. В МШ различают два основных подхода – метрический и неметрический. При *метрическом* шкалировании расстояния между объектами измеряются по шкале с уровнем измерения не ниже интервального, т. е. являются количественными; при *неметрическом* шкалировании расстояния определяются по ранговой шкале, т. е. являются порядковыми.

Заметим также, что многомерному шкалированию могут подвергаться не только объекты, но и переменные, используемые для их измерения. Расстояния между переменными могут вычисляться на основе коэффициентов корреляции между ними. Пример можно найти в работе Ю. Л. Качанова и Г. А. Сатарова «Базовые политические ценности населения России» (см. прил. на с. 203).

Этапы многомерного шкалирования.

1. Определение дизайна исследования, включающего выборку оцениваемых объектов, методы определения расстояний между объектами, при необходимости также выборку респондентов и набор измеряемых переменных.
2. Вычисление матрицы расстояний или подготовка переменных, которые будут использованы в алгоритмах, автоматически вычисляющих расстояния в процессе реализации МШ.
3. Выбор соответствующей исходным данным модели МШ и размерности пространственной карты.
4. Реализация метода МШ и интерпретация полученных результатов.
5. Оценка надежности и достоверности полученного численного решения.

Исходные данные для многомерного шкалирования весьма разнообразны, однако они всегда представляют собой одну или несколько матриц расстояний. Некоторые программы МШ обладают дополнительной возможностью вычислять матрицу расстояний на основе исходных переменных.

Многомерному шкалированию могут подвергаться: 1) выборка респондентов, ответивших на вопросы анкеты, или объектов (организаций, стран и т. п.), для которых измерены некоторые объективные показатели, позволяющие вычислить матрицу расстояний; 2) набор тестовых объектов, которые оценены по нескольким критериям одним или несколькими респондентами (экспертами).

В последнем случае мы сталкиваемся с проблемой, которую рассматривали по отношению к семантическому дифференциалу. Как и в СД, каждый респондент оценивает один и тот же тестовый набор объектов по нескольким критериям и для него создается персональная матрица данных, на основе которой строится индивидуальное семантическое пространство. Может также строиться агрегированное пространство, полученное в результате усреднения данных по всей выборке или некоторой подвыборке респондентов.

В отличие от СД при использовании МШ респонденты могут оценивать не отдельные свойства тестируемых объектов с помощью набора шкал, а непосредственно матрицу сходства / различий между объектами, отражающую особенности их восприятий и предпочтений. Таким образом, в МШ существуют прямые и непрямые подходы к сбору данных.

Прямые подходы основаны на суждениях респондентов о сходствах / различиях между объектами и предполагают их непосредственную оценку для каждой пары объектов с использованием шкалы Лайкерта или аналогичной шкалы с числом градаций 6, 7 и т. д. (табл. 2.21).

Таблица 2.21

Заполненный опросный бланк для оценки сходств между объектами

Пара объектов	Очень похожи	Похожи	Скорее похожи	И похожи и непохожи	Скорее непохожи	Непохожи	Очень непохожи
A и *B*	1	2	3	4	5	6	(7)
A и *C*	1	2	(3)	4	5	6	7
A и *D*	1	(2)	3	4	5	6	7
A и *E*	1	(2)	3	4	5	6	7
B и *C*	1	2	3	4	(5)	6	7
B и *D*	1	2	(3)	4	5	6	7
B и *E*	(1)	2	3	4	5	6	7
C и *D*	1	2	3	4	(5)	6	7
C и *E*	1	2	3	(4)	5	6	7
D и *E*	1	2	3	4	5	(6)	7

В результате получают $k(k-1)/2$ пар объектов (где k – количество тестовых объектов) и столько же расстояний между ними, которые удобно представить в форме матрицы (табл. 2.22). В данном случае: $k = 5$, получено $5 \times 4 / 2 = 10$ пар объектов и соответствующих расстояний.

Таблица 2.22

Матрица расстояний, полученная из табл. 2.21

Объект	*A*	*B*	*C*	*D*	*E*
A					
B	7				
C	3	5			
D	2	3	5		
E	2	1	4	6	

Непрямые подходы к сбору данных основаны на объективных и субъективных характеристиках шкалируемых объектов и предполагают их непосредственное измерение или оценивание по набору шкал (можно использовать, например, шкалы СД, Лайкерта, ранговые шкалы и т. п.). С помощью полученной матрицы данных *вычисляется* матрица расстояний между объектами, которая используется в качестве исходных данных для МШ.

Преимущество прямых методов изучения восприятий заключается в том, что исследователю не приходится определять набор характеристик для сравнения объектов, что, в свою очередь, усложняет последующую интерпретацию полученных результатов. Преимуществом непрямых методов является относительная простота интерпретации.

Таким образом, в многомерном шкалировании существует, как минимум, четыре критерия для классификации данных в зависимости от процедуры их сбора. Данные могут быть: 1) индивидуальными или усредненными; 2) субъективными или объективными; 3) метрическими или неметрическими; 4) полученными прямыми или непрямыми методами. *Индивидуальные данные* имеют место, когда набор тестовых объектов оценивается одним наблюдателем (испытуемым, экспертом и т. п.) или используются некоторые объективные показатели. *Усредненные данные* получаются в результате групповых или массовых опросов. *Субъективные данные* являются результатом оценивания объектов одним испытуемым или респондентами группового или массового опроса. *Объективные данные* не зависят от субъективных оценок, например, это могут быть результаты голосований в парламенте, статистические показатели стран или организаций, биологические характеристики экосистем. Данные являются *метрическими*, если расстояния между объектами измеряются по метрической (количественной) шкале, *неметрическими* – если они оцениваются по шкале порядка. *Прямые методы* сбора данных позволяют получить расстояния между объектами без предварительных вычислений, например, в ходе непосредственного оценивания респондентами сходства / различий между парами объектов (см. также табл. 2.20). При использовании *непрямых методов* сбора данных расстояния между объектами вычисляются с использованием значений переменных.

По форме представления в статистических программных средствах А. Д. Наследов выделяет три вида исходных данных: 1) квадратная симметричная матрица различий (расстояний); 2) квадратная асимметричная матрица различий; 3) матрица индивидуальных различий, которая представляет собой несколько сведенных воедино матриц различий одного и того же набора объектов, полученных от разных респондентов[1].

В *симметричной матрице*, в соответствии с геометрической традицией, расстояние между объектами *A* и *B* полагается равным расстоянию между объектами *B* и *A*. Она может быть вычислена на основе переменных, измеряющих некоторые объективные характеристики объектов, их усредненные оценки, ранги и т. п., или получена в результате оценки степени сходства между объектами одним респондентом (экспертом). В первом случае мы имеем классическую задачу снижения размерности, во втором – за-

[1] *Наследов А. Д.* Указ. соч. С. 299–301.

дачу изучения индивидуального семантического пространства одного респондента, как и при использовании метода СД. Симметричная матрица может быть представлена в треугольном виде.

Асимметричная матрица различий образуется в том случае, когда расстояние между объектами *А* и *В* не равно расстоянию между *В* и *А*. Это возможно при применении прямых методов сбора данных. А. Д. Наследов приводит социометрический пример, когда симпатия студента к однокурснице не обязательно означает взаимность[1]. Социологический пример находим у Н. К. Малхотры, который отмечает, что нередко Мексику воспринимают как более похожую на США, чем США на Мексику[2]. Очевидно, что асимметричная матрица может быть представлена только полностью.

Несколько матриц возникает, когда группа респондентов оценивает степень сходства / различий между парами объектов из одного и того же тестируемого набора.

Методы многомерного шкалирования. Для МШ разработано множество методов, различающихся используемыми мерами сходства / различий, алгоритмами размещения объектов в координатном пространстве, методами оценки соответствия полученного решения исходной матрице расстояний. Так, в работе А. Ю. Терехиной рассматривается 15 методов[3]. Методы МШ позволяют производить анализ данных о различиях или предпочтениях, которые измерены на метрическом, порядковом либо номинальном уровнях. Результирующие шкалы при этом также могут носить метрический, порядковый или номинальный характер.

Существуют различные классификации методов. Например, Н. К. Малхотра использует два критерия классификации: 1) природа данных – метрическая или неметрическая – и 2) уровень анализа данных – индивидуальный или агрегированный[4]. А. Ю. Терехина к этим двум критериям добавляет вид моделей, которые делятся на модели анализа предпочтений и модели анализа близостей[5].

Г. А. Сатаров выделяет три группы методов: 1) многомерное шкалирование матриц близости; 2) индивидуальное шкалирование; 3) неметрическое многомерное развертывание (НМР). Первая группа методов основана на визуализации в метрическом пространстве заранее подготовленной

---

[1] *Наследов А. Д.* Указ. соч. С. 299.

[2] *Малхотра Н. К.* Маркетинговые исследования. 4-е изд. М. ; СПб. ; Киев, 2007. С. 949.

[3] *Терехина А. Ю.* Анализ данных методами многомерного шкалирования. М., 1986.

[4] *Малхотра Н. К.* Указ. соч. С. 944.

[5] *Терехина А. Ю.* Указ. соч. С. 58.

матрицы мер близости или различий. Вторая группа базируется на оценках тестируемого набора объектов одним испытуемым по нескольким переменным и имеет много общего с МГК, однако результаты представляются не в виде матрицы нагрузок, а визуально. Эти два подхода объединяет использование *метрических* данных. Совокупность методов *неметрического* многомерного развертывания[1] предназначена для анализа данных о предпочтениях, организованных в виде набора ранжировок респондентами тестируемых объектов. В основе модели НМР лежит представление об объединенном психологическом пространстве латентных факторов, в котором могут быть одновременно представлены и тестируемые объекты, и оценивающие их респонденты. Для размещения респондентов в одном пространстве с объектами используется *модель идеальной точки*. В рамках этой модели предполагается, что респондент ранжирует набор объектов по степени их сходства с некоторым «идеальным» объектом, существующим в его представлении.

Ф. Янг и Д. Харрис[2] в качестве базовой, или «классической», модели МШ выделяют *модель евклидовых расстояний* с использованием одной симметричной матрицы расстояний. Полученное таким образом пространство называется евклидовым, является ортогональным, и в нем в качестве меры различий используется евклидово расстояние.

На случай с несколькими матрицами расстояний распространяются модификации *модели индивидуальных различий*: репликационное (повторное) МШ, взвешенное МШ, обобщенное МШ. Репликационное МШ предполагает, что все респонденты имеют общее латентное пространство восприятий. Взвешенное МШ используется, главным образом, в экспертных опросах, когда эксперты имеют неодинаковую квалификацию и их оценкам приписывается разный «вес». Обобщенное МШ предполагает создание единого пространства для всех респондентов, участвовавших в опросе. Модели индивидуальных различий интересны тем, что позволяют представлять в виде карты не только конфигурацию набора объектов, но также выборку респондентов, оценивавших характеристики этих объектов или степень их сходства / различий.

Выбор размерности пространственной карты. Выше было отмечено, что размерность полученного пространства не должна быть большой. Выбор размерности — это компромиссное решение между возможностью получить карту, наиболее точно воспроизводящую исходную матрицу расстояний, и возможностью проинтерпретировать полученные

[1] *Сатаров Г. А.* Многомерное шкалирование // Интерпретация и анализ данных в социологических исследованиях. М., 1987. С. 176—182.

[2] *Young F. W., Harris D. F.* Multidimensional Scaling // SPSS Professional Statistics. Chicago, 1994. P. 155—222.

результаты. Наиболее удобна для зрительного восприятия двумерная карта, однако в ряде случаев данная размерность не позволяет удовлетворительно проинтерпретировать полученные результаты, а иногда оказывается избыточной.

Н. К. Малхотра предлагает следующие принципы для определения необходимого количества измерений[1]:

- априорное определение размерности на основе теории или предшествующих исследований;
- интерпретируемость полученной карты;
- критерий изогнутости *графика стресса*, посредством которого измеряется надежность и достоверность полученных результатов (подробнее будет рассмотрен ниже): абсцисса точки, в которой наблюдается поворот или резкий изгиб графика, должна соответствовать оптимальной размерности, аналогично тому, как абсцисса точки, в которой наблюдается изгиб графика Каттелла (см. рис. 2.9), соответствует оптимальной размерности пространства главных компонент;
- удобство использования (по этому критерию рекомендуются двумерные карты).

Пример 2.4 (продолжение)

Для задачи с областными центрами априорно предполагалась двумерная карта как аналог географической.

Интерпретация результатов многомерного шкалирования. Как отмечалось выше, результаты МШ представляются в виде карты (чаще всего двумерной), на которой набор объектов размещен таким образом, чтобы расстояния между ними в максимально возможной степени соответствовали исходной матрице расстояний. На горизонтальной и вертикальной осях карты устанавливается определенный масштаб, соответственно каждый объект на карте получает численные координаты. Содержательная интерпретация осей пространства зависит от того, как по отношению к ним разместились конкретные объекты.

На рис. 2.15 представлены два латентных фактора, полученных методом МШ. Фактор x определяет различия между объектами A и B, с одной стороны, и объектом C – с другой; объекты D, E и F имеют по этому фактору «нейтральные» значения, близкие к 0. Фактор y аналогичным образом определяет различия между объектом D, с одной стороны, и объектами E и F – с другой. Очевидно, что содержательная интерпретация факторов x и y является трудной задачей и может быть выполнена только специалистом, хорошо знакомым с исследуемым материалом. Так, Г. А. Сатаров с соавторами в ряде работ анализировал методом МШ результаты голосо-

[1] *Малхотра Н. К.* Указ. соч. С. 945.

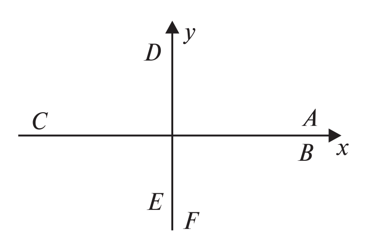


Рис. 2.15. Представление результатов многомерного шкалирования

вания в российском парламенте, где включал в выборку депутатов с яркой политической позицией[1], А. Д. Наследов в качестве примера рассматривал методы многомерного статистического анализа и студентов, которые их оценивали[2].

Координатные оси построенного пространства должны быть проинтерпретированы как факторы, определяющие различия между объектами. Другими словами, оси интерпретируются по характеристикам расположившихся вдоль них объектов. Хотя сама по себе интерпретация не входит в формальную процедуру МШ, она является хорошей проверкой того, насколько адекватно модель описывает данные и насколько решение удовлетворяет целям исследования.

Рассматривая вопросы вербальной интерпретации результатов МШ, Г. А. Сатаров указывает, что «проблема <...> возникает всякий раз, когда отсутствует априорная гипотеза о наличии определенных латентных факторов. <...> Анализируя расположение точек полученной конфигурации, мысленно или графически группируя их, проецируя на координатные оси или другие направления, исследователь должен понять и объяснить расположение точек. Опираясь на полученное решение, дополнительную информацию, свой опыт, интуицию и т. п., исследователь должен назвать свойство, объясняющее расположение точек в их проекциях на некоторое направление. <...>

При отсутствии внешних формальных критериев валидности единственный способ проверки результатов вербальной интерпретации (називания) – это обращение к речевому опыту других людей (респондентов и / или экспертов). В результате должны быть получены ответы на следующие вопросы: 1) совпадает ли понимание данного свойства исследователем с пониманием этого свойства другими людьми (при фиксированной вербализации); 2) существуют ли синонимичные названия для данного свойства; 3) существуют ли другие свойства, которые с той же степенью убедительности могут быть поставлены в соответствие той же характеристике?»[3].

Н. К. Малхотра предлагает некоторые «инструментальные» приемы, позволяющие проинтерпретировать полученные результаты наиболее точно[4].

[1] *Сатаров Г. А., Станкевич С. Б.* Расчет рейтингов законодателей: (Консерватизм и радикализм на II Съезде народных депутатов СССР) // Демократические институты в СССР: проблемы и методы исследования. М., 1991 (см. прил. на с. 208).

[2] *Наследов А. Д.* Указ. соч. С. 313–314.

[3] *Сатаров Г. А.* Указ. соч. С. 205.

[4] *Малхотра Н. К.* Указ. соч. С. 947.

1. Даже если оценки сходства (различий) объектов получены прямыми методами, о них можно собрать дополнительную информацию. Например, в маркетинговых исследованиях, если изучаются потребительские предпочтения марок, можно оценить и использовать при интерпретации их рейтинги.

2. При сборе данных прямыми методами можно также спросить респондентов, какими критериями они руководствовались, оценивая степень сходства (различий) между объектами.

3. Рекомендуется показывать респондентам пространственные карты, построенные на основе их оценок, и обсуждать с ними полученную конфигурацию точек.

4. Если существуют объективные характеристики оцениваемых объектов (например, количество километров на литр бензина для автомобиля), их можно использовать для интерпретации полученных субъективных шкал.

Оценка качества модели. Существует ряд формальных процедур, позволяющих оценить надежность и достоверность полученного пространственного решения.

Для измерения ***достоверности*** полученного решения как степени соответствия расстояний между объектами на карте исходной матрице расстояний используются два показателя: 1) квадрат коэффициента корреляции (r^2) между расстояниями по карте и в исходной матрице расстояний как показатель соответствия модели МШ исходным данным и 2) стресс Краскала как показатель неадекватности модели МШ.

Квадрат коэффициента корреляции интерпретируется как процент дисперсии исходных данных, сохраненной при многомерном шкалировании. При полном соответствии модели исходным данным $r^2 = 1$. Допустимыми считаются значения $r^2 \geq 0,6$.

Стрессом в статистике называется показатель несоответствия модели исходным данным. Стресс, вычисленный по формуле Краскала для задач МШ, представляет собой долю исходной дисперсии, которая не учтена при многомерном шкалировании. Рекомендации для интерпретации значений стресса Краскала даются в табл. 2.23.

Таблица 2.23

Интерпретация значений стресса Краскала[1]

Стресс (%)	Критерий соответствия модели
20	Плохое
10	Удовлетворительное
5	Хорошее
2,5	Отличное
0	Превосходное (полное)

[1] Источник: *Малхотра Н. К.* Указ. соч. С. 948–949.

Для оценки ***надежности*** **(*устойчивости*)** полученного решения МШ можно использовать следующие процедуры:

• исключить случайным образом из выборки некоторые объекты и сравнить позиции оставшихся объектов в старом и новом решениях;

• добавить в выборку некий искусственный объект (желательно, сконструированный таким образом, чтобы его свойства были известны[1]) или заменить им один из объектов выборки и сравнить полученные результаты с предыдущими;

• сравнить результаты, полученные на разных подвыборках или на данных, собранных в разное время.

Если полученные результаты после исключения / добавления некоторых объектов в целом не изменились, полученное решение можно считать устойчивым. Аналогично решение будет устойчивым при значительном сходстве результатов, полученных на разных подвыборках или в разное время.

Пример 2.4 (продолжение)

Вычисленное значение $R^2 = 0{,}99$, а стресс Краскала равен 1,4 %, т. е. модель весьма достоверно воспроизводит исходную матрицу расстояний.

Для проверки надежности модели из массива данных был удален г. Минск (рис. 2.16), однако это не нарушило структуры размещения оставшихся городов на карте (см. рис. 2.14, *а*).

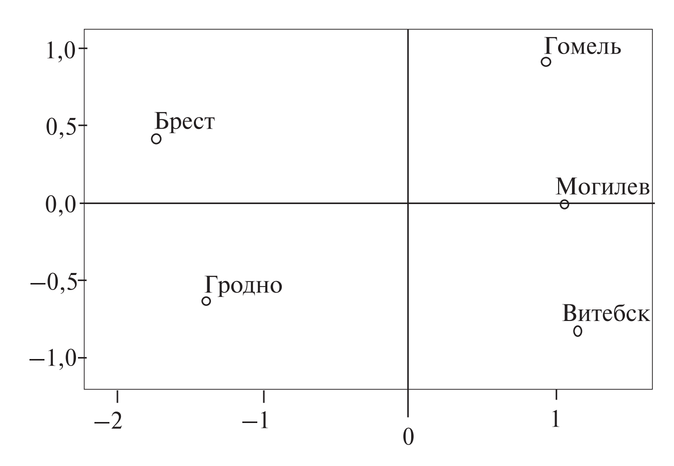


Рис. 2.16. Карта расстояний между областными центрами Беларуси после удаления из матрицы данных г. Минска

[1] Например, Г. А. Сатаров и С. Б. Станкевич в указанном сочинении «вводили» в выборку депутатов парламента искусственные объекты «идеальный консерватор» и «идеальный радикал» с соответствующими результатами голосований.

Ограничения метода многомерного шкалирования. Метод МШ отличается от всех рассмотренных выше методов по ряду характеристик.

Во-первых, выборка исследования представляет собой набор объектов, между которыми тем или иным способом определяются расстояния.

Во-вторых, использование матрицы расстояний в качестве исходных данных обусловливает специфику их получения. Расстояния могут быть получены либо прямыми методами субъективных оценок сходств или различий, либо вычислены на основе субъективных и / или объективных характеристик объектов. В первом случае нужен один или несколько экспертов, которые будут оценивать непосредственно сходства или различия между объектами. Во втором случае необходимо получить объективные характеристики объектов и / или провести опрос респондентов, который позволил бы выявить их субъективные характеристики. Переменные, измеряющие эти характеристики, должны быть пригодными для вычисления расстояний. В частности, они должны иметь одинаковый уровень измерений – дихотомический, порядковый или количественный. Причем количественных переменных не должно быть слишком много и они, по возможности, должны иметь близкий масштаб и не коррелировать друг с другом, иначе вычисленное для них евклидово расстояние не позволит построить эффективную модель.

В-третьих, интерпретация латентных факторов, представленных осями многомерной карты, осуществляется в соответствии с полученной конфигурацией объектов и их свойствами. Поэтому аналитик должен хорошо знать предметную область, иметь подробную информацию о каждом объекте, выборка объектов не может быть велика. Особенно при прямых методах сбора данных.

В-четвертых, многие содержательные задачи могут не иметь приемлемого решения, т. к. структура объектов в многомерном пространстве не всегда может быть удовлетворительно воспроизведена в пространстве небольшой размерности. В частности, А. Ю. Терехина полагает, что исследователь должен иметь некоторую уверенность в том, что система содержит сравнительно небольшое количество групп связанных объектов. Бесполезно применять методы МШ к системе, в которой все объекты связаны друг с другом и все связи имеют одинаковый порядок[1].

Анализ соответствий представляет собой разновидность метода МШ, предполагающую одновременное (на одной карте) шкалирование как объектов, так и их характеристик (рис. 2.17). Исходные данные для анализа соответствий представляют собой таблицу, строки которой соответствуют оцениваемым объектам, столбцы – приписываемым им характеристикам. На пересечении строки и столбца указывается оценка соответствия дан-

[1] *Терехина А. Ю.* Указ. соч. С. 159–160.

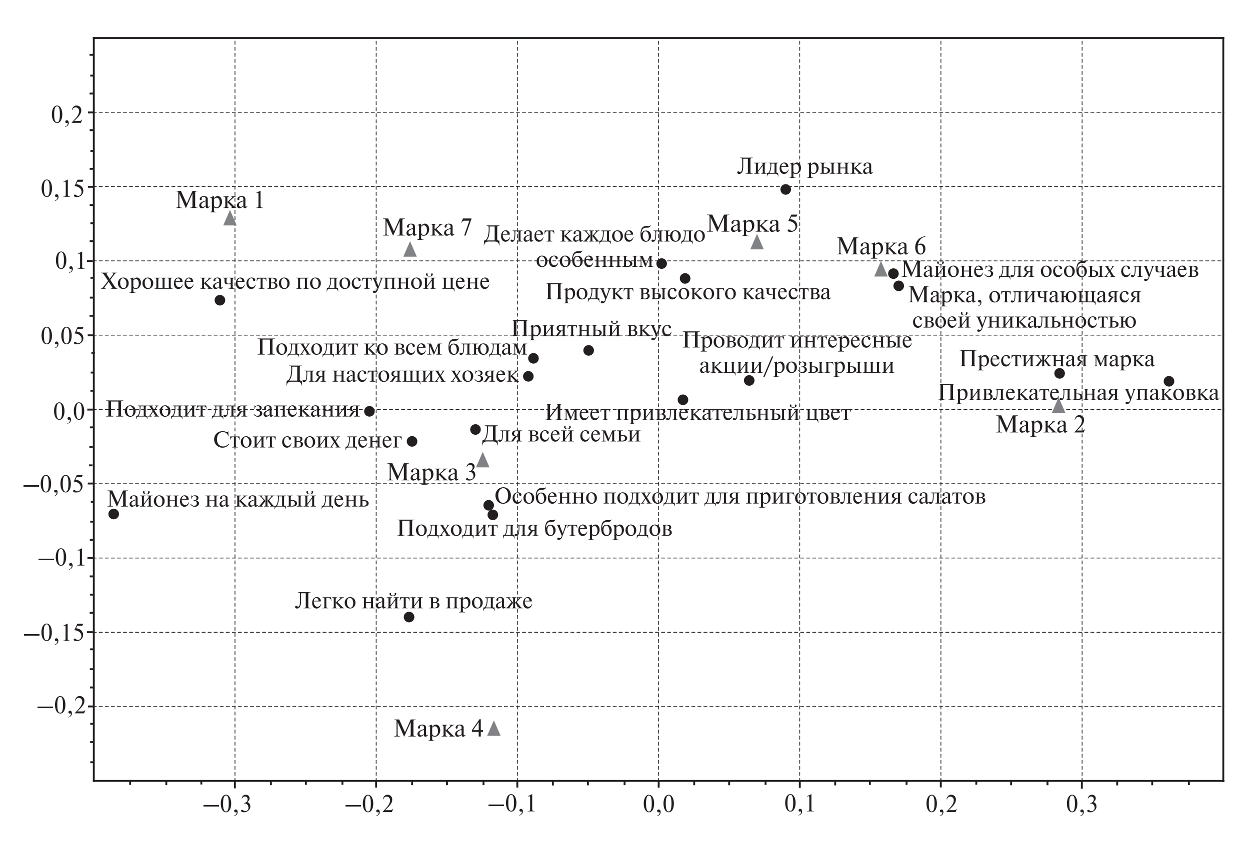


Рис. 2.17. Совместный анализ марок майонеза и их характеристик

ной характеристики определенному объекту. Оценки соответствия могут быть получены различными способами. Например, это может быть частота выбора одной характеристики из списка (в этом случае таблица соответствий является обычной таблицей сопряженности) или среднее значение степени соответствия характеристики объекту по выборке респондентов, которые оценивали соответствие по 5, 7 или 9-балльным шкалам[1].

Размещение на одной карте как объектов, так и их характеристик не только позволяет лучше понять восприятие объектов респондентами, но и облегчает интерпретацию полученных шкал.

Основное отличие анализа соответствий от многомерного шкалирования заключается в необходимости вычисления расстояний не только между характеристиками и между объектами, но также между объектами и приписываемыми им характеристиками.

Самостоятельная работа

Пользуясь рис. 2.17, проинтерпретируйте полученное семантическое пространство (горизонтальную и вертикальную оси). Охарактеризуйте марки в терминах осей пространства. Объясните, почему правый нижний («четвертый») квадрант не содержит марок.

Литература

Агапова, И. Ю. Восприятие рекламы: методика использования «репертуарных решеток» для формирования биполярных шкал семантического дифференциала / И. Ю. Агапова // Социология : 4М. 1999. № 11.

Баранова, Т. С. Психосемантические методы в социологии / Т. С. Баранова // Социология : 4М. 1993–94. № 3–4.

Бессокирная, Г. П. Факторный анализ: традиции использования и новые возможности / Г. П. Бессокирная // Социология : 4М. 2000. № 12.

Бююль, А. SPSS: Искусство обработки информации / А. Бююль, П. Цёфель. – М., 2002.

Девятко, И. Ф. Диагностическая процедура в социологии: очерк истории и теории / И. Ф. Девятко. М., 1993.

Евелькин, Г. М. Многомерное шкалирование в исследовании профессиональной социализации личности / Г. М. Евелькин. Минск, 2002.

Елисеева, И. И. Логика прикладного статистического анализа / И. И. Елисеева, В. О. Рукавишников. М., 1982.

Елисеева, И. И. Основные процедуры многомерного статистического анализа / И. И. Елисеева, Е. В. Семенова. СПб., 1993.

Захарова, Т. А. Метод совместного анализа как инструмент изучения предпочтений потребителей / Т. А. Захарова, А. Х. Кутлалиев // Социология : 4М. 2009. № 28.

---

[1] Подробнее см.: *Малхотра Н. К.* Указ. соч. С. 953.

Качанов, Ю. Л. Метаморфозы политического сознания: De omni re scibili et quibusdam alis / Ю. Л. Качанов, Г. А. Сатаров // Российский монитор: Архив современной политики. 1993. № 3.

Качанов, Ю. Л. Структура личностного образа народного депутата в сознании избирателей г. Москвы / Ю. Л. Качанов, И. В. Задорин // Демократические институты в СССР: Проблемы и методы исследования / сост. И. В. Задорин ; науч. ред. О. М. Маслова. М., 1991.

Ким, Дж.-О. Факторный анализ: Статистические методы и практические вопросы / Дж.-О. Ким, Ч. У. Мьюллер // Факторный, дискриминантный и кластерный анализ / под ред. И. С. Енюкова. М., 1989.

Крыштановский, А. О. Анализ социологических данных / А. О. Крыштановский. М., 2007.

Наследов, А. Д. SPSS: Компьютерный анализ данных в психологии и социальных науках / А. Д. Наследов. СПб., 2007.

Пажес, Ж.-П. Конфликты и общественное мнение: Новая попытка объединить социологов и математиков / Ж.-П. Пажес // Социол. исслед. 1991. № 7, 10.

Петров, В. М. Опыт применения неметрического многомерного шкалирования при изучении предпочтений молодежи в области авторской песни / В. М. Петров // Социология : 4М. 1991. № 1.

Родионова, Н. В. Семантический дифференциал : обзор литературы / Н. В. Родионова // Социология : 4М. 1996. № 7.

Сатаров, Г. А. Многомерное шкалирование / Г. А. Сатаров // Интерпретация и анализ данных в социологических исследованиях. М., 1987.

Терехина, А. Ю. Анализ данных методами многомерного шкалирования / А. Ю. Терехина. М., 1986.

Толстова, Ю. Н. Измерение в социологии / Ю. Н. Толстова. М., 1998.

Толстова, Ю. Н. Многомерное шкалирование / Ю. Н. Толстова. М., 2006.

Шафир, М. А. Анализ соответствий: представление метода / М. А. Шафир // Социология : 4М. 2009. № 28.

Глава 3

ПОСТРОЕНИЕ КЛАССИФИКАЦИЙ И ТИПОЛОГИЙ

3.1. КЛАССИФИКАЦИЯ И ТИПОЛОГИЗАЦИЯ В СОЦИАЛЬНЫХ ИССЛЕДОВАНИЯХ

Соотношение понятий «тип» и «класс». Классификация и типологизация окружающих предметов, явлений, событий является основой не только научной, но и любой человеческой деятельности. Определив тип или класс явления, человек принимает решение, основанное на предыдущем опыте, и тем самым экономит время, силы и другие ресурсы. Во многих случаях термины «классификация» и «типология» используются как синонимы, однако мы их будем различать.

Классификацией изучаемых объектов (включая респондентов, участвующих в исследовании) мы будем называть разделение их на непересекающиеся группы по каким-либо формальным критериям. Полученные при этом группы объектов являются *классами*. Классификация в традиционном понимании (назовем его четким) обладает следующим свойством: каждый объект может быть отнесен к одному и только одному классу. Другими словами, для каждого объекта существует класс, к которому он может быть отнесен, и каждый объект может быть отнесен только к одному классу. Примером реализации этого требования являются точные границы интервалов при группировке количественных переменных.

Однако в настоящее время наряду с четкой классификацией существует также концепция нечетких множеств, в рамках которой отдельные объекты не могут быть однозначно отнесены к одному из классов. Для них существует вероятность принадлежности к одному из двух или нескольких классов. В этом случае классификация состоит в том, чтобы для объекта вычислить вероятность принадлежности к каждому из классов. Причем объект не обязательно должен быть отнесен к классу, которому соответствует максимальная вероятность; он может быть также объявлен «нерасклассифицированным».

Заметим также, что термин «классификация» используется в двух значениях: как процесс разбиения объектов на непересекающиеся группы и как результат этого процесса.

Под *типом*, в соответствии с веберовской традицией[1], в социологии принято понимать идеальный объект, обладающий ярко выраженными характеристиками, отличающими его от идеальных представителей других типов или просто интересующими исследователя. Из этого определения следует, что в выборке эмпирического исследования реальный объект, совпадающий с типическим, может присутствовать, а может и не присутствовать. Процесс выделения типов называют *типологизацией*, а совокупность найденных типов — *типологией*.

Диалектика взаимоотношений понятий типа и класса — в науке вообще и в социологии в частности — заключается в том, что у исследователя есть выбор между двумя стратегиями. Первая («восходящая») стратегия состоит в построении классификации изучаемых объектов, последующем выделении типических характеристик классов и получении теоретических типов в результате интерпретации классов. Вторая («нисходящая») стратегия заключается в предварительном построении теоретической типологии, операционализации характеристик выделенных типов, разработке системы критериев для осуществления классификации, соответствующей теоретической типологии.

В любом случае тип с классом никогда полностью не совпадают. Тип трактуется как идеальная модель, гипотетический объект, на котором изучаемые явления и закономерности наблюдаются в наиболее чистом виде, а класс — как группа объектов из статистической совокупности, полученная в результате логико-математических формальных построений.

Основные подходы к построению классификаций. Какая бы стратегия ни использовалась (от классификации к типологии или от типологии к классификации), типология остается продуктом теоретического обобщения и интерпретации, а классификация — ее эмпирической базой. Поэтому при изучении анализа данных основное внимание уделяется методам классификации, в то время как задачи типологизации считаются априори решенными.

Существуют два основных подхода к решению задачи классификации. Первый подход называется *группировкой*. В результате группировки в один класс объединяются объекты, имеющие одинаковые (с точностью до интервала) значения одной или нескольких классифицирующих переменных. Наиболее часто в социальных исследованиях используются половозрастные группировки, группировки по уровню образования, доходов и др. Группировки могут осуществляться по любым измеряемым показателям:

[1] Существуют также другие определения типов (см., напр.: *Татарова Г. Г., Бессокирная Г. П.* Типологический анализ для реконструкции социальных типов работников // Социол. исслед. 2011. № 7. С. 3), но мы ограничимся данным.

электората – по кандидатам или партиям, за которые намереваются голосовать; потребителей – по предпочитаемым маркам и т. п. Группировки могут быть одномерными или многомерными. Многомерную группировку нередко называют перекрестной классификацией.

Второй подход основан на метафоре пространства переменных. В пространстве определяется мера сходства (расстояние) между объектами, и в один класс объединяются объекты, расстояния между которыми минимальны. В рамках данного подхода различают автоматическую классификацию и классификацию с обучением.

Автоматическая классификация (кластерный анализ) заключается в анализе полной матрицы расстояний между объектами, которые последовательно объединяются в кластеры.

Классификация с обучением исходит из наличия обучающей выборки, т. е. группы объектов, для которых принадлежность к классам априори известна. Основное требование к обучающей выборке состоит в том, что в ней должны быть представлены *все* классы. В некоторых случаях достаточно одного представителя класса (например, в кластерном анализе методом *k* средних), в других их должно быть несколько (например, в дискриминантном анализе). При классификации с обучением в многомерном пространстве переменных определяются (нередко вычисляются) геометрические центры будущих классов – *центроиды*, затем вычисляется матрица расстояний между центроидами, определенными по обучающей выборке, и объектами, подлежащими классификации. Как правило, объект относится к тому классу, расстояние до центроида которого минимально. Однако иногда вместо расстояний (или наряду с ними) используются вероятности принадлежности к каждому из классов. Тогда объект относят к классу, вероятность быть отнесенным к которому максимальна.

Автоматическая классификация является реализацией «восходящей» исследовательской стратегии «от класса к типу»; классификация с обучением – «нисходящей» стратегии «от типа к классу». Группировка может быть использована в каждой из стратегий, в зависимости от того, как исследователь выбирает для нее группирующие переменные.

В задачах классификации всегда существуют два вида переменных: 1) переменные, являющиеся критериями классификации или используемые при вычислении расстояний (будем называть их *классификационными* переменными); 2) переменные, которые не участвуют в выделении классов, но в дальнейшем могут использоваться при изучении различий между классами и их интерпретации (будем называть их *контрольными* переменными).

Любая классификационная схема основывается на общем фундаментальном принципе, состоящем из двух положений: 1) в один класс объединяются объекты, сходные между собой в некотором смысле; 2) степень

сходства между объектами внутри класса должна быть больше, чем степень сходства между объектами, относящимися к разным классам. Другими словами, первое положение отражает максимальную однородность внутри классов; второе – максимальные различия между классами[1].

Группировка. Использование статистических методов анализа данных основано на предположении об однородности исследуемой совокупности объектов. Это требование однако не сводится к ограничению объекта изучения. Любая реальная статистическая совокупность практически всегда внутренне дифференцирована и представляет собой своеобразный комплекс отличных друг от друга объектов или явлений. Это делает задачу разделения исследуемой совокупности на однородные группы актуальной при анализе систем объектов любой природы.

Для группировок используются переменные с любым уровнем измерения: номинальные, порядковые и количественные (последние должны быть сгруппированы в интервалы).

Группировка объектов с использованием одной, существенной с точки зрения целей исследования, переменной может рассматриваться как *одномерная классификация*. Такой переменной может быть какая-либо социально-демографическая характеристика (пол, возраст, образование, место жительства и др.) или некий аспект поведения респондентов (наличие или отсутствие вредных привычек, посещение или непосещение театров или дискотек, ценностные ориентации и др.). Одномерная группировка является, по сути, одномерным частотным распределением. В «чистом» виде она представляет структуру выборочной совокупности в разрезе одной классифицирующей переменной. Одномерная группировка позволяет также сравнивать между собой группы, образованные значениями классифицирующей переменной, по любым другим измеряемым показателям. Например, сравнивать возрастные группы по средним доходам или по проценту пользователей сети Интернет.

Двумерная группировка (с использованием двух классифицирующих переменных) получила название перекрестной классификации. С точки зрения статистического анализа данных двумерная группировка представляет собой таблицу сопряженности. Подобно одномерной группировке, она может использоваться для сравнительного анализа групп, выделенных по значениям двух классифицирующих переменных, описания двумерной структуры статистической совокупности, а также изучения корреляционных и причинных связей между переменными. В двух последних видах задач используется одинаковое представление данных, различия состоят в выборе классифицирующих переменных. Для изучения структуры чаще выбираются социально-демографические или другие контролируе-

[1] *Елисеева И. И., Рукавишников В. О.* Группировка, корреляция, распознавание образов. М., 1977. С. 9–10.

мые переменные. При изучении связей переменные могут быть как социально-демографическими, так и измеряющими различные поведенческие аспекты, ценности, мотивы и т. п.

Многомерная группировка по существу достаточно близка к двумерной, однако представление результатов перекрестной классификации по трем или более переменным сталкивается с большими техническими трудностями. Наиболее простым и распространенным является подход, получивший название лингвистической классификации. Данные, представленные в соответствии с этим подходом, удобно использовать при решении сравнительных и структурных задач. Аналогичным образом представляются данные при анализе связей в многомерных (многовходовых) таблицах сопряженности методом логлинейного анализа, который мы не будем здесь рассматривать подробно[1].

Лингвистическая классификация – это, строго говоря, не отдельный метод, а техника, которая позволяет идентифицировать классы, построенные в результате многомерной перекрестной классификации, при помощи уникальной аббревиатуры, состоящей из последовательности букв (цифр, знаков), соответствующих значениям классифицирующих переменных. В первую очередь определяется порядок классификационных переменных. Затем каждому значению каждой переменной присваивается собственное обозначение. Каждому объекту из выборки ставится в соответствие аббревиатура, состоящая из последовательности кодов значений переменных, после чего подсчитывается количество объектов с одинаковой аббревиатурой.

Заметим, что при группировке внутригрупповая однородность и межгрупповые различия для классифицирующих переменных обеспечиваются автоматически. Что касается внутригрупповой однородности и межгрупповых различий по контрольным переменным, их наличие или отсутствие определяется в ходе решения задач сравнительного анализа с использованием средних арифметических, дисперсий, распределений частот и др.

Пример 3.1. Археологические данные[2]

Данные о раскопках представляют собой информацию о 25 древних захоронениях (табл. 3.1). Три переменные (возраст, пол, статус захороненного) являются классификационными. По ним осуществлялась лингвистическая классификация: первая буква означает возраст захороненного (*C* – ребенок, *T* – подросток, *A* – взрослый); вторая – пол (*M* – мужской, *F* – женский); третья – социальный статус (*E* – элитарный, *N* – неэлитарный).

[1] См., напр.: *Аптон Г.* Анализ таблиц сопряженности. М.,1982 ; *Наследов А. Д.* Указ. соч. Гл. 25.

[2] Источник: *Олдендерфер М. С., Блэшфилд Р. К.* Кластерный анализ // Факторный, дискриминантный и кластерный анализ / под ред. И. С. Енюкова. М., 1989. С. 210.

Таблица 3.1

Археологические данные

№ п/п	Класс	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
1	*CMN*	1	0	0	1	0	0	0	0
2	*CMN*	0	0	0	1	0	0	0	0
3	*CME*	1	0	0	1	0	0	1	1
4	*CFN*	1	0	1	0	0	0	0	0
5	*CFE*	0	0	1	0	0	0	1	0
6	*CFE*	1	0	1	0	0	0	1	0
7	*TMN*	1	1	0	1	0	0	0	0
8	*TMN*	0	1	0	1	1	0	0	0
9	*TMN*	1	0	0	1	1	0	0	0
10	*TMN*	1	1	0	1	1	0	0	0
11	*TME*	1	1	0	1	1	0	1	1
12	*TFN*	0	0	0	0	1	0	0	0
13	*TFN*	1	0	0	0	1	0	0	0
14	*TFE*	1	0	0	0	1	0	1	0
15	*AMN*	1	1	0	1	1	0	0	0
16	*AMN*	0	1	0	1	1	0	0	0
17	*AMN*	1	1	0	1	0	0	0	0
18	*AME*	1	1	0	1	1	0	1	1
19	*AME*	1	0	0	1	0	0	1	0
20	*AFN*	0	0	0	0	0	1	0	0
21	*AFN*	1	0	0	0	0	1	0	0
22	*AFN*	0	0	0	0	1	1	0	0
23	*AFN*	1	0	0	0	0	0	0	0
24	*AFE*	1	0	0	0	1	1	1	0
25	*AFE*	1	0	0	0	1	1	1	1

Остальные 8 переменных являются контрольными; они имеют дихотомический уровень измерения и фиксируют наличие или отсутствие в захоронениях определенных предметов: x_1 – местная керамика; x_2 – наконечники стрел; x_3 – обломки браслетов; x_4 – обработанные камни; x_5 – костяные иглы; x_6 – костяные шилья; x_7 – привозная керамика; x_8 – металлические изделия.

Сравнительный анализ предметов в захоронениях в зависимости от возраста, пола и статуса (лингвистическая классификация) представлен в табл. 3.2. Во втором столбце (*n*) указано количество захоронений соответствующего класса; в столбцах, содержащих переменные x_1, x_2 ... x_8, – доля захоронений данного класса, в которых обнаружен соответствующий предмет.

Таблица 3.2

Распределение древних захоронений по классам

Класс	n	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
CMN	2	0,5	0	0	1	0	0	0	0
CME	1	1	0	0	1	0	0	1	1
CFN	1	1	0	1	0	0	0	0	0
CFE	2	0,5	0	1	0	0	0	1	0
TMN	4	0,75	0,75	0	1	0,75	0	0	0
TME	1	1	1	0	1	1	0	1	1
TFN	2	0,5	0	0	0	1	0	0	0
TFE	2	1	0	0	0	1	0	1	0
AMN	3	0,67	1	0	1	1	0	0	0
AME	2	1	0,5	0	1	0,5	0	1	0,5
AFN	4	0,5	0	0	0	0,25	0,75	0	0
AFE	2	1	0	0	0	1	1	1	0,5

3.2. КЛАСТЕРНЫЙ АНАЛИЗ

Методы кластерного анализа: общая характеристика. Кластерный анализ, или автоматическая классификация, – общее название большой группы методов, использующих для классификации объектов расстояния между ними в многомерном пространстве переменных. Термином «кластер» принято называть группу объектов, расстояния между которыми значительно меньше, чем расстояния до объектов из других кластеров. Кластерный анализ появился в начале 1960-х гг. как часть структурной лингвистики при «обучении» компьютеров человеческому языку, состоявшем в распознавании букв алфавитов (и других знаков) независимо от начертания. Отсюда другие, теперь редко употребляемые, названия данной группы методов: «таксономия»[1] и «распознавание образов». Эти методы не существовали в «классической» статистике и изначально были ориентированы на использование компьютерных алгоритмов.

Тогда же появились термины «классификация с обучением» и «классификация без обучения». Классификация с обучением предполагает образование кластеров «вокруг» эталонных образцов, с которыми сравниваются

[1] Термин «таксономия» означает учение о принципах и практике классификации и систематизации; впервые предложен в 1813 г. швейцарским ботаником О. Декандолем, занимавшимся классификацией растений.

классифицируемые объекты. При классификации без обучения кластеры образуются из «похожих» друг на друга объектов без использования какой-либо априорной информации о количестве и характере классов. Отсутствие требования априорной информации объясняет популярность методов кластерного анализа в социальных науках, т. к. в большом количестве классификационных задач такая информация просто отсутствует.

Из определения кластера как группы объектов, расстояния между которыми значительно меньше, чем расстояния до объектов из других кластеров, следует, что наилучшие результаты кластерный анализ может принести в ситуации хорошо различимых множеств объектов (рис. 3.1), характерной для различения и распознавания символов в структурной лингвистике.

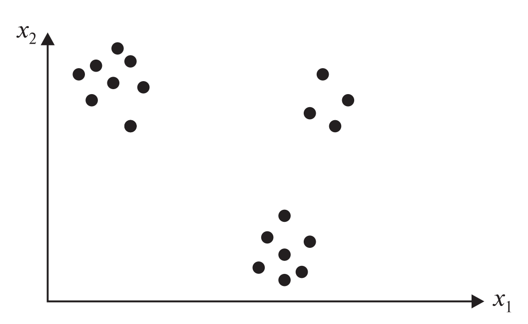


Рис. 3.1. Хорошо различимые группы объектов в пространстве двух переменных

Методы кластерного анализа (КА) являются непараметрическими методами классификации, т. е. они не предполагают ограничений на распределение исходных переменных и поэтому могут применяться к данным любого уровня измерения (количественным и качественным). Существует большое число моделей КА, которые различаются алгоритмами классификации и применяемыми мерами расстояний[1]. Некоторые варианты КА могут использоваться для классификации не только объектов, но и переменных[2], характеризующих объекты, — в этом случае может также решаться задача снижения размерности.

Таким образом, *целью* КА является исследование структуры выборочной совокупности объектов или / и набора переменных. Методы КА могут применяться как в эксплораторных, так и в конфирматорных исследованиях. *Эксплораторный* (*разведывательный*) *КА* позволяет построить классификацию объектов из выборки и по результатам интерпре-

[1] *Олдендерфер М. С., Блэшфилд Р. К.* Указ. соч.

[2] *Качанов Ю. Л., Сатаров Г. А.* Базовые политические ценности населения России // Информационные материалы. М., 1991.

тации полученных классов выделить типы объектов, встречающиеся в генеральной совокупности. *Конфирматорный* (*подтверждающий*) *КА* может использоваться для проверки гипотез о существовании типов объектов, выделенных теоретически, их распространенности, принадлежности некоторых объектов к определенному типу, наличии связей между переменными и др.

В качестве исходных данных для эксплораторного КА служит матрица расстояний между объектами, для конфирматорного КА – матрица расстояний между объектами и эталонными образцами.

Методы кластерного анализа можно классифицировать по следующим основаниям:

1) объект кластерного анализа (совокупность объектов или набор переменных);

2) кластеризация с обучением или без обучения;

3) виды используемых мер сходства / различий (расстояний) между объектами (см. разд. 2.1);

4) виды алгоритмов: иерархические (агломеративные и дивизимные) и неиерархические (оптимизация заданного критерия качества, поиск сгущений и др.).

Этапы кластерного анализа. Принято выделять следующие этапы КА:

1) формирование выборки объектов для кластеризации;

2) определение множества переменных, по которым будут оцениваться расстояния между объектами;

3) выбор адекватной меры сходства / различий между парами объектов, вычисление матрицы расстояний;

4) применение одного из методов КА для создания групп сходных объектов;

5) определение количества кластеров и их интерпретация;

6) проверка достоверности результатов кластерного решения.

Выборка объектов для кластеризации, в первую очередь, должна быть гетерогенной, иными словами, входящие в нее объекты должны отличаться разнообразием (в идеале, происходить из разных генеральных совокупностей[1]). Данные из примера 3.1 соответствуют этому требованию, т. к. захоронения различаются по возрасту, полу и социальному статусу и можно ожидать, что могильники также будут содержать разные культурные предметы. Аналогично в примере 1.1 представлены генеральные совокупно-

[1] Этому требованию идеально соответствуют задачи структурной лингвистики, где все возможные начертания буквы «а» составляют одну генеральную совокупность, а начертания группы «б» – другую.

сти стран Западной, Центральной и Восточной Европы, различающихся по уровню благосостояния и благополучию демографической ситуации.

При кластеризации без обучения существуют также ограничения на объем выборки. Во-первых, результаты представляются графически, а слишком большая дендрограмма[1] сложна для интерпретации. Во-вторых, интерпретация кластеров имеет нечто общее с интерпретацией осей многомерного шкалирования, а именно: во многих случаях кластеры можно проинтерпретировать, только обладая определенной дополнительной информацией об объектах из выборки (что делает затруднительным мысленные манипуляции большим количеством объектов). В-третьих, многие методы КА не отличаются устойчивостью, результаты могут измениться, даже если изменить порядок объектов в выборке. Вероятность таких эффектов увеличивается пропорционально объему выборки.

Выбор переменных для кластеризации объектов. Как уже отмечалось, методы КА являются непараметрическими, т. е. не предполагают нормальности распределения исходных переменных, что позволяет широко применять их в социальных науках. Выбор переменных для КА, тем не менее, не так прост и имеет, как минимум, два аспекта. Переменные должны, во-первых, наилучшим образом отражать сходство и различия между объектами из выборки и, во-вторых, быть пригодными для вычисления хотя бы одной из мер сходства или различий.

Способность переменной описывать сходство и / или различия между объектами называется информативностью. Иногда информативность переменной можно определить априори, в других случаях она может проявиться (или не проявиться) только на этапе проверки достоверности кластерного решения.

Пример 3.1 (продолжение)

Для анализа информативности контрольных переменных воспользуемся табл. 3.2.

Например, переменная x_1 не является информативной, т. к. местная керамика встречается во всех классах захоронений и, следовательно, по ее присутствию установить принадлежность захоронения к определенному классу нельзя. Переменная x_2 обладает ограниченной информативностью: наконечники стрел встречаются в захоронениях взрослых и подростков мужского пола, но не всегда. Следовательно, наличие данного предмета позволяет отнести захоронение к одному из этих классов, однако его отсутствие не свидетельствует о возможности уверенно причислить захоронение к классу женских. Переменная x_7 является высокоинформативной, т. к. ее наличие позволяет четко выделить группу захоронений детей женского пола (встречается во всех таких захоронениях и только в них).

[1] Дендрограмма – график, демонстрирующий последовательность объединения объектов в кластеры.

Выбор меры сходства или различий между объектами из выборки также не является очевидным. Таких мер достаточно много[1], однако использование каждой из них ограничено определенной конфигурацией пространства переменных и уровнем их измерения. Оптимальным является набор переменных с одинаковым уровнем измерения, когда все переменные дихотомические, порядковые или количественные. Кроме того, для количественных переменных обычно используется евклидово расстояние, следовательно, они не должны коррелировать между собой. Если переменных много, вряд ли удастся избежать эффекта корреляции между ними, поэтому во многих случаях рекомендуется предварительно применить один из методов снижения размерности и построить относительно небольшое ортогональное пространство (этот подход иногда называют «кластеры на факторах»[2]).

Реализация методов кластерного анализа. Помимо принципиальной неустойчивости, о которой говорилось выше, методы КА обладают, по сравнению с остальными методами многомерной статистики, и другими существенными недостатками. Многие из них представляют собой скорее эвристические алгоритмы и не имеют достаточного статистического обоснования. Разные методы нередко порождают для одних и тех же данных разные решения. Поэтому для выбора метода КА не существует достаточно обоснованных рекомендаций. Если исследование носит разведывательный характер, основным критерием адекватности кластерного решения является *интерпретируемость*. Методы КА нередко приходится последовательно «перебирать» до тех пор, пока не будет получено удовлетворительное с точки зрения интерпретации решение либо пока исследователь не убедится, что задача классификации для данной выборки с данным набором переменных решения не имеет.

Здесь мы рассмотрим более подробно иерархический агломеративный кластерный анализ, пользующийся, по оценке М. С. Олдендерфера и Р. К. Блэшфилда[3], наибольшей популярностью среди исследователей, а также неиерархический метод *k* средних.

Иерархические агломеративные алгоритмы предполагают пошаговое объединение объектов, начиная с наиболее близких друг к другу, до тех пор, пока все объекты не будут составлять один кластер.

На первом шаге каждый объект считается отдельным кластером. Анализируется полная матрица расстояний между объектами, и объединяются два или несколько объектов, расстояния между которыми минимальны, т. е. объединяются объекты, наиболее близкие друг к другу.

---

[1] *Олдендерфер М. С., Блэшфилд Р. К.* Указ. соч. С. 149–165.

[2] *Галицкая Е. Г., Галицкий Е. Б.* Кластеры на факторах: как избежать распространенных ошибок? // Социология : 4М. 2006. № 22.

[3] *Олдендерфер М. С., Блэшфилд Р. К.* Указ. соч. С. 167.

На каждом из последующих шагов происходит одно из трех событий:

1) объединяется следующая по степени близости пара объектов;

2) объект объединяется с кластером, к которому он наиболее близок;

3) объединяются два или несколько наиболее близких кластеров.

Процесс продолжается до тех пор, пока все объекты не войдут в один кластер.

Иерархические агломеративные методы различаются по правилам объединения кластеров, в основе которых лежит способ измерения расстояний между кластерами[1]:

- среднее расстояние между объектами из двух кластеров (метод межгрупповой связи);
- среднее расстояние между объектами из двух кластеров с учетом расстояний внутри кластеров (метод внутригрупповой связи);
- расстояние между ближайшими объектами из двух кластеров (метод ближнего соседа);
- расстояние между самыми далекими объектами из двух кластеров (метод дальнего соседа);
- расстояние между центрами кластеров (центроидный метод);
- расстояние между центрами кластеров с учетом их численности (медианный метод);
- метод Уорда (Ward), позволяющий оптимизировать минимальную дисперсию между классами.

Подробное описание этих методов можно найти в специальной литературе[2].

Общие черты иерархических агломеративных методов КА:

1) на каждом шаге матрица расстояний пересчитывается (вычисляются расстояния между созданными ранее кластерами, а также объектами, не вошедшими к началу данного этапа ни в один кластер) и последовательно объединяются наиболее близкие кластеры и объекты;

2) последовательность объединений визуально представляется в виде графика, получившего название кластерного дерева (древовидной диаграммы, дендрограммы);

3) для полной кластеризации требуется максимум $n-1$ шаг, где n – объем выборки;

4) в результате получаются непересекающиеся кластеры, которые одновременно являются вложенными в том смысле, что каждый кластер может рассматриваться как элемент другого, более широкого кластера на более низком уровне сходства.

---

[1] *Крыштановский А. О.* Анализ социологических данных. М., 2007. С. 207.

[2] *Олдендерфер М. С., Блэшфилд Р. К.* Указ. соч. С. 165–191.

Иерархические дивизимные методы КА (логическая противоположность агломеративных методов) предполагают пошаговое «расслоение» выборки на все более мелкие кластеры до тех пор, пока каждый объект не будет составлять отдельный класс. В начале процедуры кластеризации все объекты принадлежат к одному кластеру, который затем как бы «разрезается» на последовательно уменьшающиеся «ломтики».

Пример 3.1 (продолжение)

На рис. 3.2 представлена дендрограмма классификации захоронений иерархическим методом Уорда. Для классификации использованы переменные $x_2 \dots x_7$. Переменные x_1 и x_8 не включены в анализ как низкоинформативные (см. табл. 3.2). Объекты на дендрограмме обозначены номерами, а также аббревиатурой, полученной в результате лингвистической классификации (см. табл. 3.1).

Условные расстояния между кластерами

Объект	Номер
AFE	24
AFE	25
TFE	14
TFN	12
TFN	13
AFN	22
CFE	5
CFE	6
AFN	20
AFN	21
CFN	4
AFN	23
TME	11
AME	18
AMN	15
AMN	16
TMN	8
TMN	10
CME	3
AME	19
TMN	7
AMN	17
CMN	1
CMN	2
TMN	9

Шкала: 0, 5, 10, 15, 20, 25 (+---------+---------+---------+---------+---------+); метка *1*

Рис. 3.2. Дендрограмма иерархического кластерного анализа

Кластерный анализ методом k средних – метод классификации с обучением, предполагающий «быстрое» формирование кластеров вокруг предварительно заданных центроидов. Успех зависит главным образом от того, насколько удачно априорно определены количество кластеров и их центроиды. Вычисляется матрица расстояний от каждого центроида до каждого из объектов, подлежащих классификации. Процедура классификации контролируется соотношением дисперсий внутри и вне кластера и корректируется таким образом, чтобы данное отношение достигало максимума.

Использование этого метода возможно и тогда, когда оценки центроидов отсутствуют и имеется только предположение о количестве кластеров k. В этом случае k первых объектов в файле данных «назначаются» центроидами, вследствие чего результаты становятся значительно менее предсказуемыми, особенно при большом объеме выборки, который в данном случае не ограничен, т. к. результаты представляются не в виде дендрограммы, а в виде списков объектов, отнесенных к каждому кластеру.

Определение числа кластеров и их интерпретация. Результаты неиерархической кластеризации не требуют графического выражения и представляются в виде списка (таблицы) объектов, отнесенных к каждому кластеру. Результаты иерархической кластеризации представляются в виде специальных графиков – дендрограмм. Это самый распространенный способ представления результатов иерархических методов КА. Дендрограммы отражают последовательность объединения кластеров от первого до последнего шага. Каждый шаг, на котором объединялось два или несколько объектов и / или кластеров, представляется «ветвью» этого дерева. В целом дендрограмма графически изображает иерархическую структуру, порожденную матрицей расстояний и правилом объединения объектов в кластеры.

Процесс объединения объектов и кластеров на дендрограмме представлен в направлении слева направо (см. рис. 3.2). Процесс заканчивается объединением всех объектов в один общий кластер. Над кластерным деревом расположена шкала условных расстояний между объединяемыми объектами и кластерами (в данном случае расстояние между двумя большими кластерами, объединенными на последнем шаге, составляет 25 условных единиц).

Двигаясь по дендрограмме справа налево и «разрезая» последовательно каждую «ветвь», получают все более мелкие кластеры для интерпретации.

Пример 3.1 (продолжение)

Для определения числа интерпретируемых кластеров рассмотрим дендрограмму в обратном направлении – справа налево (см. рис. 3.2). Если «разрезать» вертикальную скобку на расстоянии $d = 25$, можно получить два больших кластера (место «разреза» отмечено косой чертой).

Эти два кластера (объекты с номерами с 24 по 23 и с 11 по 9) легко интерпретируются как «женский» (верхний) и «мужской» (нижний). Однако выделение более мелких кластеров нецелесообразно, т. к. ни один из них не является гомогенным.

Пример 1.1 (продолжение)

Кластерный анализ европейских стран осуществлен по методике «кластеры на факторах» (рис. 3.3).

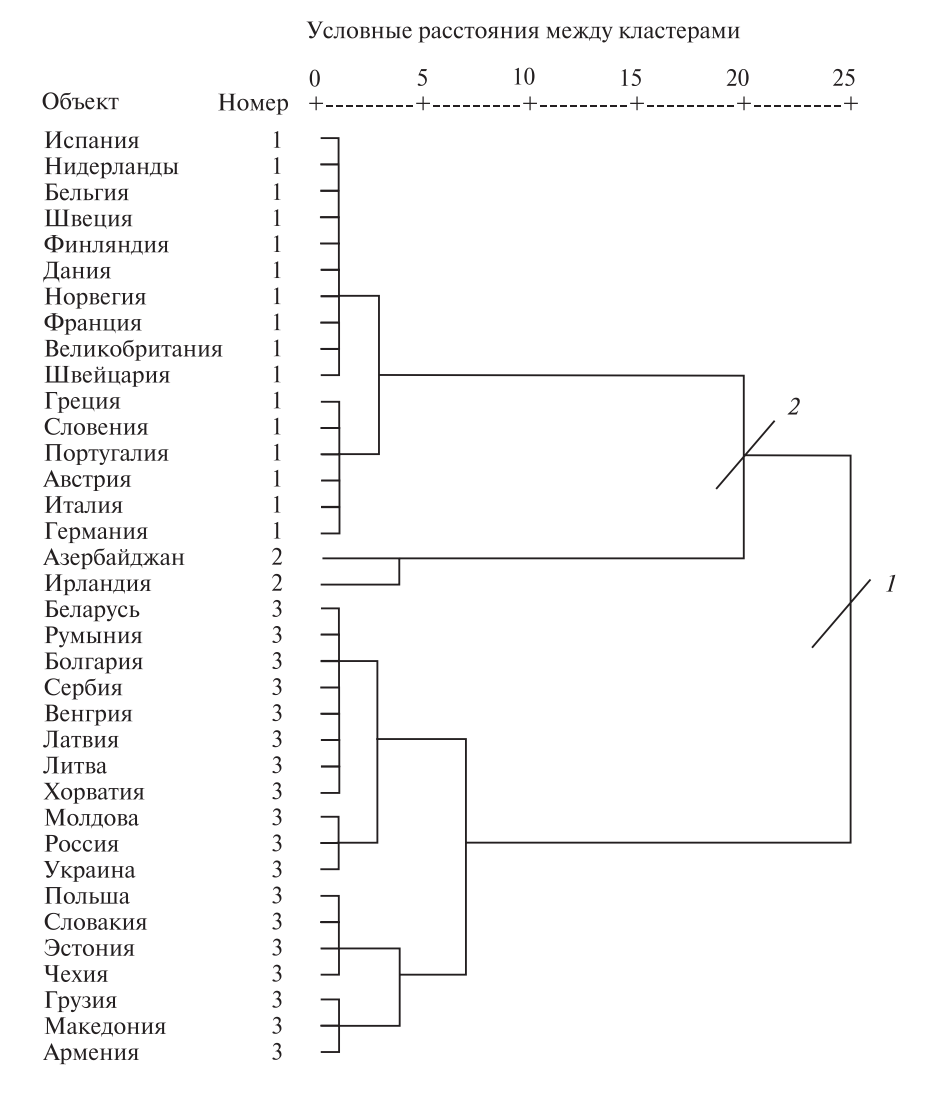


Рис. 3.3. Кластерный анализ европейских стран (2 клатера)

В разд. 2.4 был проведен факторный анализ нескольких показателей социально-демографического и экономического развития стран Европы. Два вычисленных фактора — «уровень благосостояния» и «демографическая ситуация» (табл. 2.14) — стали переменными, на которых осуществлена кластеризация методом Уорда с использованием евклидова расстояния. «Разрежем» дендрограмму на несколько кластеров, продвигаясь справа налево («разрезы» пронумерованы).

После двух «разрезов» получим три кластера (рис. 3.3): 1) страны Западной Европы (верхний кластер, от Испании до Германии); 2) Азербайджан и Ирландия (в центре); 3) страны Центральной и Восточной Европы (нижний кластер, от Беларуси до Армении). Чтобы лучше понять различия между кластерами и проинтерпретировать кластер, состоящий из двух стран (Азербайджана и Ирландии), построим диаграмму рассеяния (рис. 3.4), на которой названия стран заменим номерами кластеров, к которым они принадлежат (см. столбец номеров кластеров на рис. 3.3).

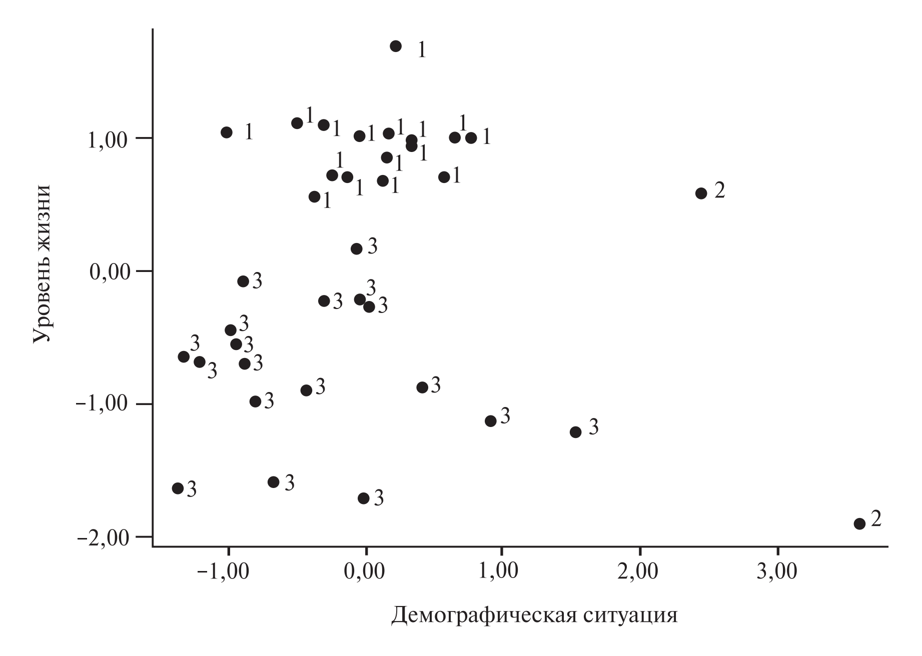


Рис. 3.4. Распределение трех кластеров европейских стран в пространстве двух факторов

На графике видно, что западноевропейские страны и страны Центральной и Восточной Европы имеют сопоставимое состояние демографической ситуации и различаются по уровню жизни. Ирландия и Азербайджан объединены в один кластер на основании благоприятной демографической ситуации (вне зависимости от уровня жизни).

Продолжая движение по дендрограмме справа налево, можно разукрупнять интерпретируемые кластеры и увеличивать их количество (рис. 3.5).

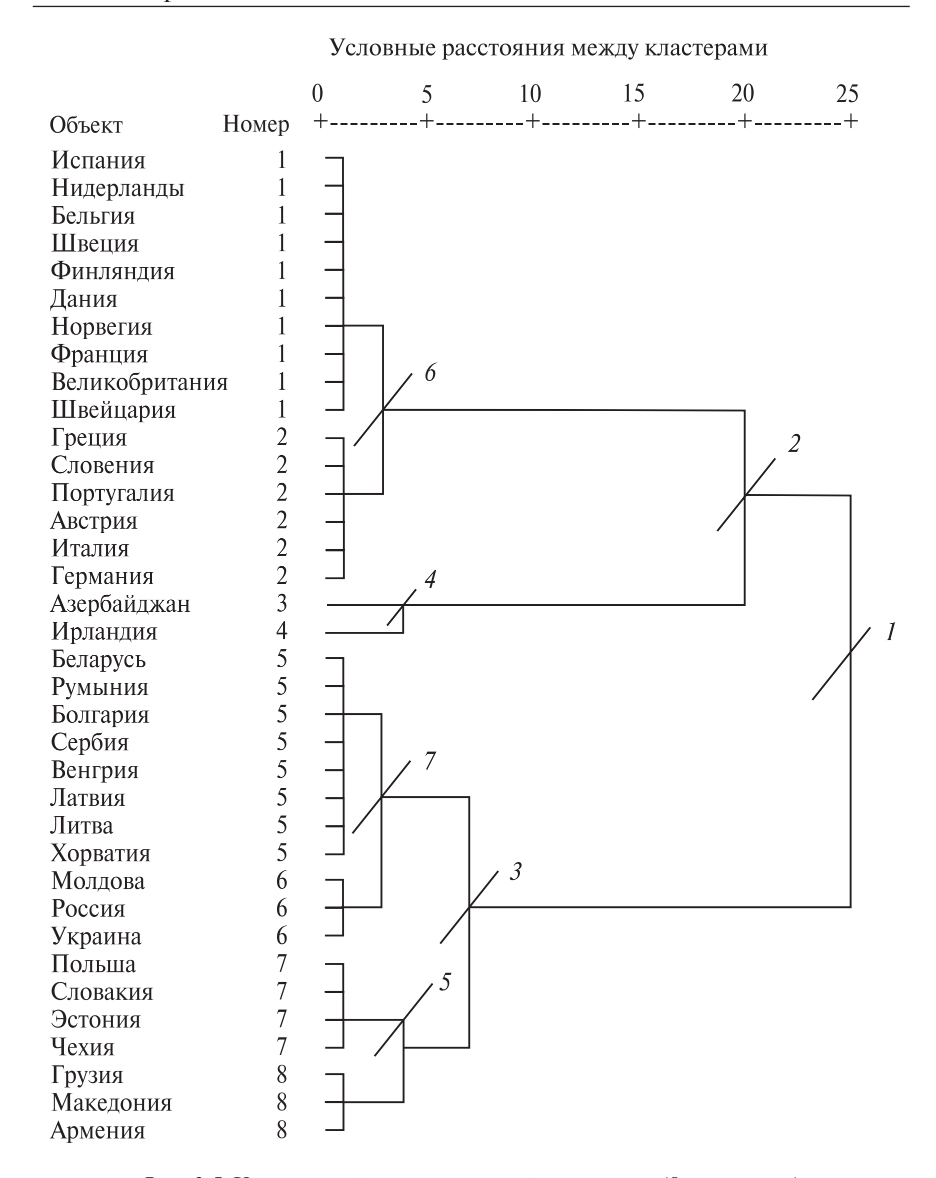


Рис. 3.5. Кластерный анализ европейских стран (8 кластеров)

В предельном случае, дойдя до нижнего уровня образования кластеров, можно выделить 8 кластеров, два из которых состоят из одной страны каждый (Азербайджан и Ирландия). Диаграмма на рис. 3.6 позволяет локализовать и проинтерпретировать каждый кластер.

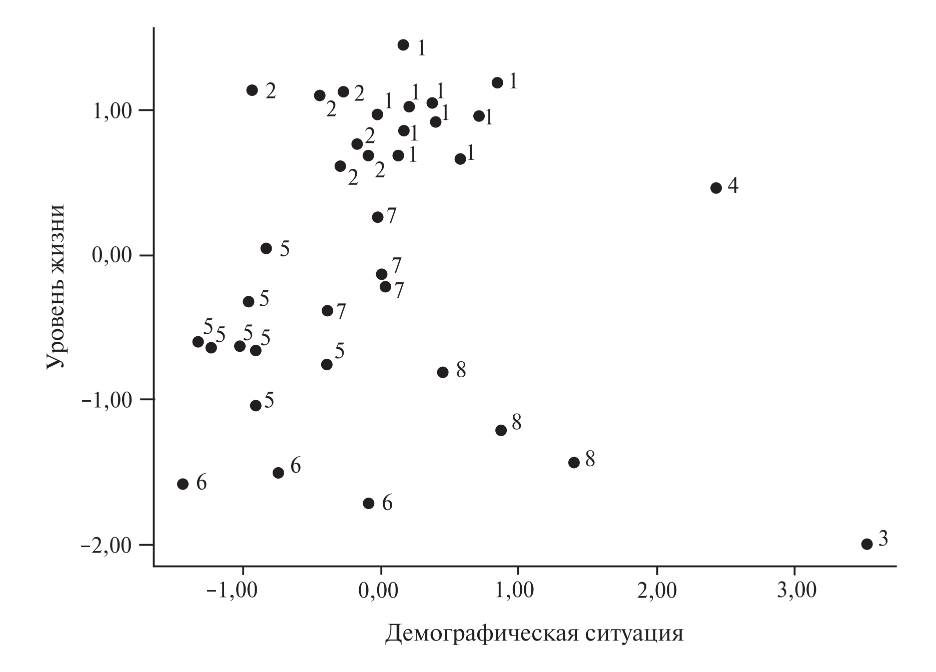


Рис. 3.6. Распределение восьми кластеров в пространстве двух факторов

Выбор количества интерпретируемых кластеров и собственно интерпретация – в значительной мере эвристическая процедура. Связанные с этим решения зависят, в частности, от постановки задачи. Так, анализ трех кластеров в последнем примере (см. рис. 3.3 и 3.4) дает возможность описать различия между западноевропейскими странами, с одной стороны, и центрально- и восточноевропейскими странами, с другой стороны. Анализ более мелких восьми кластеров (см. рис. 3.5 и 3.6) позволяет внутри двух больших кластеров выделить мелкие гомогенные группы стран и описать различия между ними.

Важным критерием является интерпретируемость кластеров. Например, в кластерном решении на рис. 3.2 не имеет смысла разукрупнять два больших кластера, т. к. отсутствует возможность интерпретировать более мелкие кластеры.

Проверка достоверности результатов кластерного решения является заключительным этапом КА. На этой стадии чаще всего применяются следующие подходы.

1. Сопоставление полученных кластеров с теоретическими представлениями исследователя (если они есть). Так, в примере 3.1 можно было предположить, что кластеры будут соответствовать половозрастной и статусной структуре древней общности. Однако самый успешный из большо-

го количества попыток вариант, полученный методом Уорда (см. рис. 3.2), позволил обнаружить существенные различия главным образом по полу.

2. Сравнение кластеров по внешним показателям, не использовавшимся в КА (например, по социально-демографическим характеристикам и др.), в том числе с применением тестов статистической значимости обнаруженных различий.

Более изысканный способ верификации результатов КА внешними переменными применили С. В. Сивуха и М. Х. Титма[1], использовавшие для этой цели метод логистической регрессии.

3. Сопоставление конечных результатов с результатами, полученными другими методами. Для знакомства с этим подходом можно рекомендовать работу Г. А. Сатарова и С. Б. Станкевича[2], где результаты КА депутатов верифицировались результатами многомерного шкалирования.

Пример 1.1 (продолжение)

Для двух больших кластеров (см. рис. 3.3) – Западной Европы (1) и Центральной и Восточной Европы (3) – внешним критерием является их географическая связанность.

Можно также сравнить средние значения *исходных* переменных (в то время как классификация проводилась в пространстве двух факторов). Из табл. 3.3 видно, что при схожей возрастной структуре (медианный возраст) и одинаковом уровне рождаемости, но при более чем двукратной разности в ВВП, показатели, связанные со смертностью (включая ожидаемую продолжительность жизни женщин и мужчин), значительно более благоприятны в Западной Европе, чем в Центральной и Восточной, что также согласуется с теорией.

Таблица 3.3

Сравнение кластеров по средним значениям переменных

Кластер	Медианный возраст	Рождаемость	Смертность	Продолжит. жизни мужчин	Продолжит. жизни женщин	Детская смертность	ВВП
1	40,787	10,9835	9,2129	77,833	83,093	4,3733	23282,27
3	38,242	11,0176	12,0056	69,084	77,653	13,1011	10060,47

Кластеризация переменных. Как отмечалось в начале данного раздела, кластерному анализу может подвергаться не только выборка объектов, но и набор переменных (рис. 3.7). Кластерный анализ может проводиться с целью изучения структуры набора переменных или проверки предположений о такой структуре (например, в психометрическом тесте). По резуль-

[1] *Сивуха С. В., Титма М.* Социальные детерминанты самооценки успеха // Социальное расслоение возрастной когорты. М., 1997.

[2] *Сатаров Г. А., Станкевич С. Б.* Расчет рейтингов законодателей: (Консерватизм и радикализм на II Съезде народных депутатов СССР) // Демократические институты в СССР: проблемы и методы исследований. М., 1991.

татам кластерного анализа может быть осуществлено снижение размерности пространства переменных. Однако, в отличие от факторного анализа или многомерного шкалирования, снижение размерности не осуществляется в рамках самого метода кластерного анализа, а должно выполняться дополнительно, с помощью каких-либо вспомогательных процедур (например, аналогичных семантическому дифференциалу).



Рис. 3.7. Кластерный анализ переменных: предпочтения музыкальных жанров

Самостоятельная работа

1. Разработайте измерительную методику на основе дендрограммы жанров музыки (см. рис. 3.7).
2. Проанализируйте кластерное дерево бюджетов времени различных социальных групп[1], используя табл. 3.4 и рис. 3.8.
3. Проинтерпретируйте полученные кластеры.

В табл. 3.4 представлены усредненные бюджеты времени за 100 суток по социальным группам, выделенным по следующим показателям: пол, занятость, маритальный статус, регион проживания. Обозначения групп приведены после таблицы.

Таблица 3.4

Бюджеты времени

Группа	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
HAUS	610	140	60	10	120	95	115	60	175	315
FAUS	475	90	250	30	140	120	100	775	115	430
FNUS	10	0	495	110	170	110	130	785	160	430

[1] Источник: *Жамбю М.* Иерархический кластерный анализ и соответствия. М., 1988. С. 23.

Окончание табл. 3.4

Группа	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
HMUS	615	140	65	10	115	90	115	765	180	305
FMUS	179	29	421	87	161	112	119	776	143	373
HCUS	585	115	50	0	150	105	100	760	150	385
FCUS	482	94	196	18	141	130	96	775	132	336
HAWE	653	100	95	7	57	85	150	808	115	330
FAWE	511	70	307	30	80	95	142	816	87	262
FNWE	20	7	568	87	112	90	180	843	125	386
HMWE	665	97	97	10	52	85	152	807	122	320
FMWE	168	22	528	69	102	83	174	824	119	311
HCWE	643	105	72	0	62	77	140	813	100	388
FCWE	429	34	262	14	92	97	147	849	84	392
HAES	650	142	122	22	76	94	100	764	96	334
FAES	578	106	338	42	106	94	92	752	64	228
FNES	24	8	594	72	158	92	128	840	86	398
HMES	652	133	134	22	68	94	102	763	122	310
FMES	436	79	433	60	119	90	107	772	73	231
HCES	627	148	68	0	88	92	86	770	58	463
FCES	434	86	297	21	129	102	94	799	58	380
HAYO	630	140	120	15	85	90	105	760	70	365
FAYO	560	105	375	45	90	90	95	745	60	235
FNYO	10	10	710	55	145	85	130	815	60	380
HMYO	650	145	112	15	85	90	105	760	80	358
FMYO	260	52	576	59	116	85	117	775	65	295
HCYO	615	125	95	0	115	90	85	760	40	475
FCYO	433	89	318	23	112	96	102	774	45	408

Пpмечание. *H* – мужской
F – женский
A – работает
N – не работает
M – женат
C – одинок
US – США
WE – Западная Европа
ES – Восточная Европа
YO – Югославия

x_1 – работа
x_2 – транспорт
x_3 – домашнее хозяйство
x_4 – дети
x_5 – покупки
x_6 – личное время
x_7 – пища
x_8 – сон
x_9 – телевидение
x_{10} – досуг

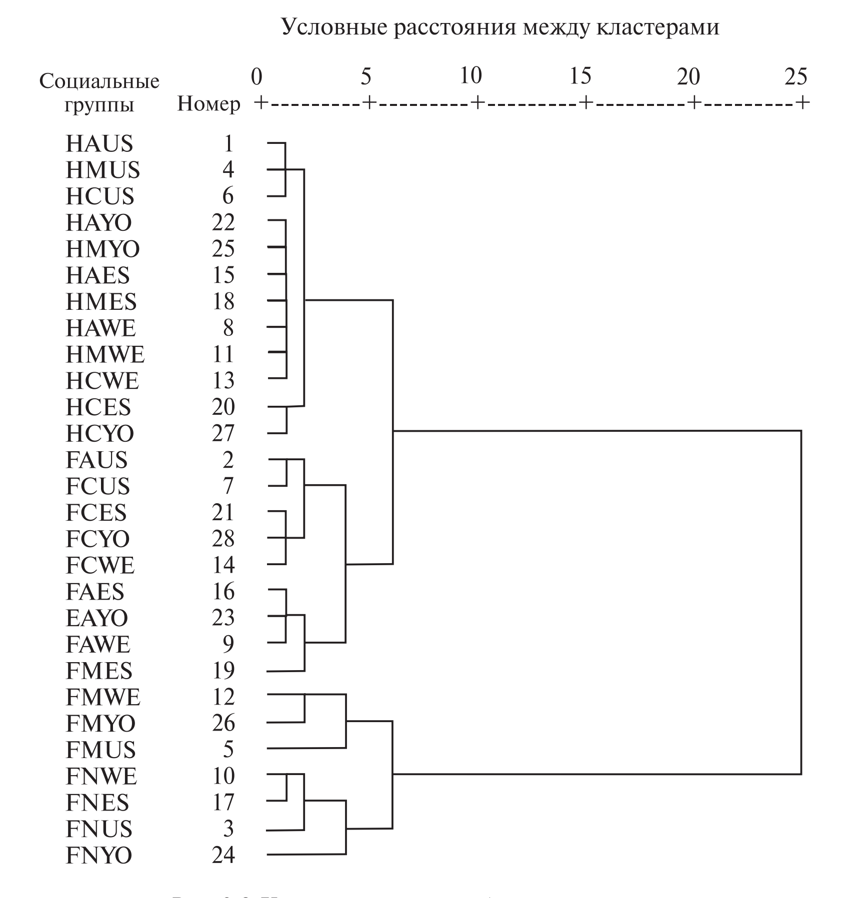


Рис. 3.8. Кластерный анализ бюджетов времени

Результаты кластерного анализа данных о бюджете времени социальных групп представлены на рис. 3.8.

3.3. КЛАССИФИКАЦИЯ С ОБУЧЕНИЕМ: ЛИНЕЙНЫЙ ДИСКРИМИНАНТНЫЙ АНАЛИЗ

Общая характеристика методов дискриминантного анализа. Методы дискриминантного анализа, наряду с некоторыми методами кластерного анализа, относятся к числу методов классификации с обучением. Они могут применяться, если известно точное количество классов и для каждого класса имеются «образцы» представителей, образующие обучающую выборку. «Образцы» могут быть двух видов: во-первых, это объекты из вы-

борки, для которых принадлежность к классу известна (например, в маркетинге это могут быть клиенты, купившие или не купившие некоторый продукт); во-вторых, это теоретические «объекты», сформированные на основе операционализации представлений исследователя об идеальных типах (например, тип «идеального» деятеля определенного политического направления, применяемый в исследованиях политических элит[1]).

Наличие обучающей выборки является общей особенностью всех кластерных и дискриминантных методов классификации с обучением. Общим является также использование для классификации расстояний между центроидами классов в обучающей выборке и объектами, подлежащими классификации. Центроидом называется геометрический центр класса в пространстве переменных, координаты которого вычисляются как среднее арифметическое координат объектов, входящих в данный класс. В дискриминантном анализе координаты центроида вычисляются по обучающей выборке. В кластерном анализе допускается, чтобы класс был представлен в обучающей выборке только одним объектом, который и становится центроидом.

Основным различием между кластерным и дискриминантным анализом является то, что методы кластерного анализа используют для классификации исключительно расстояния («вклад» отдельных переменных в классификацию не учитывается) и результаты могут быть проинтерпретированы только при наличии у исследователя дополнительной (формализованной или неформализованной) информации о классифицируемых объектах; в то время как дискриминантные методы позволяют определить «вклад» в классификацию переменных, характеризующих классифицируемые объекты, и использовать их для интерпретации различий между классами. Соответственно классифицирующие переменные, называемые в дискриминантном анализе *дискриминантными переменными*, должны быть количественными (шкалы не ниже интервальных). Еще одно отличие заключается в том, что дискриминантный анализ позволяет оценить вероятность принадлежности классифицируемого объекта к каждому из классов, а не просто приписать его к одному из них.

Обучающая выборка в дискриминантном анализе должна состоять не менее чем из двух представителей каждого класса.

Этапы дискриминантного анализа. Из вышеизложенного следует, что применение дискриминантного анализа предполагает предварительное решение вопросов о количестве и характере классов, формировании обучающей выборки, выборе или построении пространства дискриминантных переменных, отвечающих определенным требованиям. В целом можно выделить четыре этапа дискриминантного анализа.

[1] *Сатаров Г. А., Станкевич С. Б.* Указ. соч.

На *первом этапе* необходимо определить количество и характер классов, сформировать обучающую выборку из «ярких» представителей каждого класса (не менее двух).

На *втором этапе* формируется пространство дискриминантных переменных, обладающих хорошей дифференцирующей (разделительной) способностью или информативностью. Информативной является переменная, значения которой для центроидов классов максимально различаются между собой.

На рис. 3.9 переменная x_1 является информативной (обладает высокой дифференцирующей способностью), т. к. для центроидов двух классов C_1 и C_2 ее значения $x_1(C_1)$ и $x_1(C_2)$ весьма различаются. Переменная x_2 не является информативной, т. к. разность ее значений для центроидов $x_2(C_1)$ и $x_2(C_2)$ невелика.

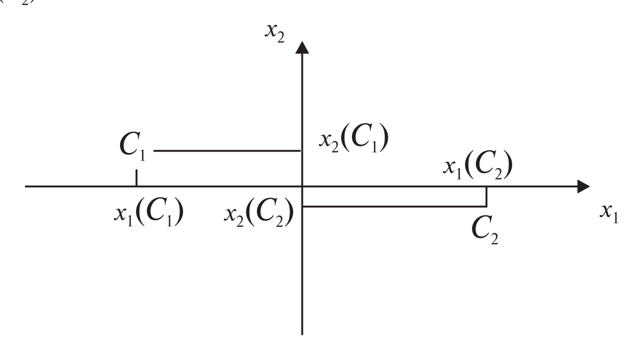


Рис. 3.9. Информативная (x_1) и неинформативная (x_2) переменные

При наличии большого числа переменных, имеющих высокую разделительную способность, для построения дискриминантного пространства могут применяться методы снижения размерности, например факторный анализ или многомерное шкалирование. Дискриминантные переменные могут быть также получены в результате концептуального агрегирования большого числа переменных-индикаторов, как это было сделано, например, в исследовании Б. А. Бардес, процитированном в работе У. Р. Клекки[1].

Заметим, что первые два этапа дискриминантного анализа могут иногда выполняться параллельно (как в названном исследовании Бардес), и даже в обратном порядке, как в работе О. В. Терещенко[2], где группы респондентов с характерными особенностями выбирались в пространстве, предварительно построенном методом факторного анализа.

[1] *Клекка У. Р.* Дискриминантный анализ // Факторный, дискриминантный и кластерный анализ. М., 1989.

[2] *Терещенко О. В.* Классификация социологических объектов методами многомерного статистического анализа // Вестн. БГУ. Сер. 3. 1994. № 2.

На *третьем этапе* производится классификация объектов, не вошедших в обучающую выборку, по степени их «похожести» на центроиды выделенных классов. Однако алгоритмы классификации в дискриминантном анализе, как будет показано ниже, существенно отличаются от алгоритмов кластерного анализа.

Четвертым, заключительным, этапом построения классификации является проверка качества полученного разделения на классы. Для этого используется переклассификация объектов из обучающей выборки по построенной модели аналогично тому, как это делается в логистической регрессии (см. разд. 1.3). Могут также применяться контрольные переменные и математический аппарат проверки гипотез. В целом, методы дискриминантного анализа, в отличие от методов кластерного анализа, не являются сугубо «механистическими», а сочетают строгие математические подходы с эвристическими.

Требования к модели дискриминантного анализа. Дискриминантный анализ (ДА) относится к методам многомерной статистики, позволяющим изучать различия между несколькими группами (классами) объектов по набору переменных и классифицировать новые объекты, для которых класс первоначально не был определен. Метод ДА широко применяется в социологии, психологии, экономике, политологии, медицине, биологии и других науках.

Целью ДА является решение двух взаимосвязанных задач: интерпретации и классификации. В случае *интерпретации* необходимо определить: можно ли, используя данный набор переменных, отличить один класс от другого; насколько хорошо эти переменные позволяют различать классы; и какие из них наиболее информативны. В случае *классификации* необходимо решить, к какому классу принадлежит классифицируемый объект (для объектов из обучающей выборки — принадлежит ли объект к данному классу). Переменные, применяемые для того, чтобы отличать один класс от другого, и образующие пространство, в котором происходит классификация, называются дискриминантными (разделяющими) переменными.

У. Р. Клекка формулирует ряд требований к модели дискриминантного анализа[1]. В модели должно быть:

1) не менее двух классов ($g \geq 2$);

2) не менее двух объектов в каждом классе: $n_i \geq 2$ $\left(i = \overline{1,g}\right)$, составляющих обучающую выборку;

3) любое число дискриминантных переменных p, не превосходящее общее число объектов в обучающей выборке n за вычетом двух ($0 < p < (n - 2)$);

4) измерение дискриминантных переменных по интервальной шкале или шкале отношений;

[1] *Клекка У. Р.* Указ. соч. С. 84.

5) линейная независимость дискриминантных переменных (дискриминантные переменные не должны коррелировать друг с другом);

6) приблизительное равенство между ковариационными матрицами для каждого класса;

7) многомерная нормальность распределения дискриминантных переменных для каждого класса.

Требования равенства ковариационных матриц и многомерной нормальности распределений лежат в основе методов ДА, однако его процедуры достаточно устойчивы по отношению к незначительным нарушениям этих требований и не всегда требуют их выполнения[1].

Канонические дискриминантные функции. Пространство канонических дискриминантных функций является основным инструментом как интерпретации различий между классами, так и классификации объектов. Его можно рассматривать как аналог пространства главных компонент, с той разницей, что назначение главных компонент – максимально эффективно описать различия между *всеми* объектами из выборки, а канонических дискриминантных функций – между *центроидами классов*. Заметим также, что канонические дискриминантные функции, в отличие от главных компонент, могут представляться как в стандартизованном, так и в нестандартизованном виде.

Каноническая дискриминантная функция является линейной комбинацией дискриминантных переменных и удовлетворяет определенным условиям. Она имеет следующее математическое представление:

$$f_i = u_{i,0} + u_{i,1}x_1 + u_{i,2}x_2 + \ldots + u_{i,p}x_p = u_{i,0} + \sum_{j=1}^{p} u_{i,j}x_j, \qquad (3.1)$$

где f_i – каноническая дискриминантная функция с номером i ($i = \overline{1,k}$); x_j – дискриминантная переменная с номером j $\left(j = \overline{1,p}\right)$; $u_{i,j}$ – коэффициенты, обеспечивающие выполнение требуемых условий.

Коэффициенты $u_{1,j}$ для первой функции выбираются таким образом, чтобы ее значения для центроидов классов как можно больше отличались друг от друга (имели максимальную дисперсию). Коэффициенты $u_{2,j}$ для второй функции выбираются так же, но при этом налагается дополнительное условие ортогональности: вторая функция не должна коррелировать с первой. Аналогично выбираются коэффициенты для третьей и последующих канонических дискриминантных функций.

Максимальное число канонических дискриминантных функций, которое можно получить описанным способом, равно числу классов без единицы или числу дискриминантных переменных, в зависимости от того, какая из этих величин меньше ($k = \min(g - 1, p)$). Например, для разде-

[1] Подробнее см.: *Клекка У. Р.* Указ. соч. С. 130–132.

ления двух центроидов достаточно одной функции, независимо от количества дискриминантных переменных, для разделения трех центроидов – максимум две функции и т. д. С другой стороны, количество канонических дискриминантных функций не может превышать числа исходных дискриминантных переменных.

Использование всех дискриминантных функций, которые могут быть получены, не всегда эффективно. Для определения их оптимального количества применяется тот же подход, что и в методе главных компонент: вычисляются собственные значения матрицы корреляции исходных дискриминантных переменных, соответствующие дискриминантным функциям, и доля общей дисперсии, приходящаяся на каждую из них.

Центр (начало координат) пространства дискриминантных функций находится в «главном» центроиде – точке геометрического пространства, координаты которой вычислены как средние арифметические дискриминантных переменных по всей обучающей выборке.

Пример 3.2. Маркетинг газонокосилок[1]

На рис. 3.10 представлена обучающая выборка: по 12 домохозяйств, купивших и не купивших газонокосилку в зависимости от размера участка и уровня доходов. Задача состоит в построении модели, которая позволила бы по размеру участка и уровню доходов предсказать, будет ли совершена покупка.

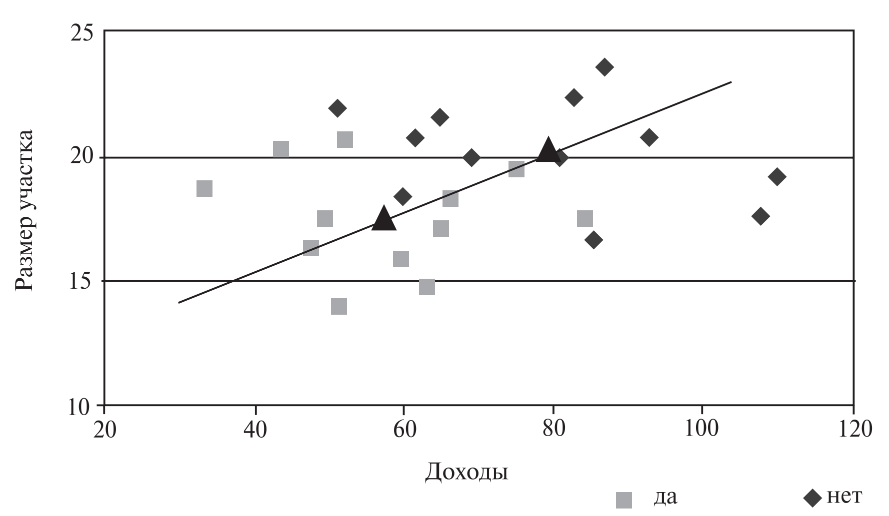


Рис. 3.10. Покупка газонокосилки в зависимости от размера участка и уровня доходов

[1] Источник: *Johnson R. A., Wichern D. W.* Applied Multivariate statistics. London, 1992. P. 496.

Обе переменные – размер участка и уровень доходов – полезны для предсказания покупки или отказа от нее, однако ни одна из них по отдельности не позволяет удовлетворительно разделить эти два класса.

Центроиды двух классов представлены на рисунке черными треугольниками. Координаты центроида класса домохозяйств, купивших газонокосилку, – (57,33; 17,63), координаты центроида класса домохозяйств, отказавшихся от покупки, – (79,47; 20,27). Прямая, соединяющая два центроида, и есть каноническая дискриминантная функция *f*, в данном случае единственная, т. к. разделяются только два класса.

Начало координат пространства, состоящего из единственной канонической дискриминантной функции *f*, находится в точке с координатами (68,40; 18,95).

Канонические дискриминантные функции могут быть представлены не только в нестандартизированном (формула 3.1), но также в стандартизированном и структурном виде. Заметим, что в стандартизированных и структурных функциях отсутствует свободный член уравнения $u_{i,0}$.

Нестандартизированные коэффициенты, аналогично нестандартизированным коэффициентам регрессии, используются для вычисления значений канонических дискриминантных функций (дискриминантных значений) для классифицируемых объектов. Таким образом, происходит переход от исходного пространства дискриминантных переменных к пространству канонических дискриминантных функций.

В пространстве нестандартизированных дискриминантных функций осуществляется классификация новых объектов и тестирование разделяющей способности модели посредством переклассификации объектов из обучающей выборки. Для этого необходимо вычислить: 1) дискриминантные значения для каждого центроида; 2) дискриминантные значения для каждого объекта; 3) расстояния в дискриминантном пространстве от каждого объекта до каждого центроида. Объект относится к классу, до центроида которого у него минимальное расстояние.

Стандартизированные коэффициенты, аналогично стандартизированным коэффициентам регрессии, позволяют сравнить *вклады* дискриминантных переменных в значения дискриминантных функций.

Структурные коэффициенты представляют собой коэффициенты корреляции между каноническими дискриминантными функциями и дискриминантными переменными. Таким образом, они являются аналогами нагрузок в факторном анализе и используются для интерпретации канонических дискриминантных функций. Наряду с общими вычисляются также внутригрупповые структурные коэффициенты, отражающие взаимосвязь канонических функций и дискриминантных переменных в пределах отдельных групп (классов).

Структурные и стандартизированные коэффициенты рекомендуется рассматривать вместе, т. к. в случае высокой корреляции между двумя

дискриминантными переменными это позволяет определить вклад каждой переменной в дискриминантное значение более точно.

Все виды коэффициентов представляются в форме матриц, аналогичных матрице стандартизированных и нестандартизированных коэффициентов регрессии: строки соответствуют дискриминантным переменным, а столбцы – каноническим дискриминантным функциям.

Пример 3.2 (продолжение)

В нашем примере только два класса и, соответственно, только одна каноническая дискриминантная функция, поэтому для экономии места мы свели три матрицы в одну (табл. 3.5).

Таблица 3.5

Коэффициенты канонических дискриминантных функций

Дискриминантная переменная	Каноническая дискриминантная функция		
	нестандартизированный коэффициент	стандартизированный коэффициент	структурный коэффициент
Доходы	0,049	0,808	0,641
Площадь участка	0,380	0,786	0,614
Constant	-10,522	–	–

Вычислим значения канонической дискриминантной функции для центроидов. Центроид группы домохозяйств, купивших газонокосилку («да»), имеет следующие координаты: средние доходы – 79,47 (тыс. долл. в год); средняя площадь участка – 20,27 (тыс. кв. футов); значение канонической дискриминантной функции для него $f(\textit{да}) = 0{,}049 \times 79{,}47 + 0{,}38 \times 20{,}27 - 10{,}522 = 1{,}038$. Аналогично для группы домохозяйств, не купивших газонокосилку («нет»): $f(\textit{нет}) = -1{,}038$.

Представим себе домохозяйство *A*, имеющее доходы 70 тыс. долл. в год и участок площадью 18 тыс. кв. футов: $f(A) = 0{,}049 \times 70 + 0{,}38 \times 18 - 10{,}522 = 0{,}248$. Расстояние в одномерном каноническом дискриминантном пространстве от домохозяйства *A* до центроида группы «да» $d(A, \text{да}) = \left|1{,}038 - (-0{,}248)\right| = 1{,}286$ [1], до центроида группы «нет» аналогично: $d(A, \text{нет}) = \left|-1{,}038 - (-0{,}248)\right| = 0{,}79$. Следовательно, можно предсказать, что домохозяйство *A* газонокосилку скорее всего не купит.

Классификационные функции. Помимо хорошо обоснованного подхода к классификации в пространстве канонических дискриминантных функций, в дискриминантном анализе применяются также подходы, основанные на использовании линейных классификационных функций и исходных дискриминантных переменных.

---

[1] Формула для вычисления расстояния обусловлена одномерностью канонического дискриминантного пространства; в пространстве большей размерности необходимо применять многомерное евклидово расстояние.

В первом случае применяется набор из g (по количеству классов) специальных функций, предназначенных исключительно для классификации. Каждая из них соответствует одному классу и представляет собой линейную комбинацию дискриминантных переменных:

$$h_i = b_{i,0} + b_{i,1}x_1 + b_{i,2}x_2 + \ldots + b_{i,p}x_p = b_{i,0} + \sum_{j=1}^{p} b_{i,j}x_j, \qquad (3.2)$$

где i — номер класса ($i = \overline{1, g}$); x_j — дискриминантные переменные ($j = \overline{1, p}$); $b_{i,j}$ — связывающие их коэффициенты.

Объект относится к классу, которому соответствует максимальное из вычисленных значений классификационных функций (max (h_i), $i = \overline{1, g}$).

Коэффициенты простых классификационных функций представляются в виде матрицы, аналогичной матрице коэффициентов мультиноминальной логистической регрессии, по одному столбцу на каждый класс (табл. 3.6).

Таблица 3.6

Классификационные функции для двух классов

Дискриминантная переменная	Купил газонокосилку	
	нет	да
Доходы	0,331	0,432
Площадь участка	4,693	5,482
Constant	-51,558	-73,398

Во втором случае процедура аналогична кластерному анализу с обучением: в пространстве дискриминантных функций вычисляются расстояния от объекта до каждого из центроидов, и объект относится к ближайшему классу. При этом используется расстояние Махаланобиса D^2, представляющее собой обобщенное евклидово расстояние для неортогональных пространств (т. к. к исходным дискриминантным переменным требование ортогональности не предъявляется).

При соблюдении требования многомерной нормальности дискриминантных переменных внутри групп (см. с. 134) свойства расстояния Махаланобиса таковы, что позволяют вычислить также апостериорную вероятность[1] принадлежности объекта к каждому из классов (сумма апостериорных вероятностей для всех классов равна единице). Как правило,

[1] В методах классификации с обучением различают априорную и апостериорную вероятность принадлежности объекта к каждому из классов. Априорная вероятность задается исследователем исходя из теоретических представлений или вычисляется как доля класса в обучающей выборке; апостериорная вероятность определяется как доля соответствующего класса после классификации с использованием построенной модели. Подробнее см.: *Клекка У. Р.* Указ. соч. С. 114–117.

минимальному расстоянию до центроида соответствует максимальная вероятность принадлежности к соответствующему классу. Однако довольно часто встречается ситуация, когда объект находится примерно на одинаковом расстоянии от двух или более центроидов и, соответственно, имеет близкие по величине вероятности быть отнесенным к каждому из них. Знание этих вероятностей позволяет исследователю в спорных случаях самостоятельно принять решение, к какому классу следует отнести объект, что существенно отличается от возможностей автоматической классификации.

Использование расстояния Махаланобиса позволяет также учитывать при классификации априорные вероятности классов, задаваемые исследователем.

Иногда в сложных случаях целесообразно отказаться от классификации «спорных» объектов и объявить их неподдающимися классификации.

Пример 3.2 (продолжение)

В табл. 3.7 показано, что модель, включающая две дискриминантные переменные (доходы и площадь участка), из 24 объектов, входящих в обучающую выборку, правильно классифицирует 21 объект и неправильно три (№ 1, 13 и 17). Объект № 2 находится практически на одинаковом расстоянии от обоих центроидов и имеет одинаковые вероятности для отнесения к двум классам.

Таблица 3.7

Вероятность принадлежности объекта к каждому из двух классов

Номер случая	Актуальная группа	Предсказанная группа			Вторая по близости группа			Значение канонической функции
		Группа	Вероятность	D^2 Махаланобиса	Группа	Вероятность	D^2 Махаланобиса	
1	**Да**	**Нет**	0,783	0,177	**Да**	0,217	2,738	-0,617
2	Да	Да	0,507	1,051	Нет	0,493	1,104	0,013
3	Да	Да	0,849	0,042	Нет	0,151	3,497	0,832
4	Да	Да	0,682	0,449	Нет	0,318	1,976	0,368
5	Да	Да	0,996	2,664	Нет	0,004	13,747	2,670
6	Да	Да	0,988	1,169	Нет	0,012	9,966	2,119
7	Да	Да	0,949	0,138	Нет	0,051	5,988	1,409
8	Да	Да	0,985	0,945	Нет	0,015	9,289	2,010
9	Да	Да	0,709	0,372	Нет	0,291	2,148	0,428

Окончание табл. 3.7

Номер случая	Актуальная группа	Предсказанная группа			Вторая по близости группа			Значение канонической функции
		Группа	Вероятность	D^2 Махаланобиса	Группа	Вероятность	D^2 Махаланобиса	
10	Да	Да	0,981	0,738	Нет	0,019	8,614	1,897
11	Да	Да	0,657	0,524	Нет	0,343	1,828	0,314
12	Да	Да	0,891	0,001	Нет	0,109	4,196	1,011
13	**Нет**	**Да**	0,764	0,221	**Нет**	0,236	2,576	0,567
14	Нет	Нет	0,548	0,892	Да	0,452	1,279	-0,093
15	Нет	Нет	0,851	0,039	Да	0,149	3,526	-0,840
16	Нет	Нет	0,802	0,133	Да	0,198	2,925	-0,673
17	**Нет**	**Да**	0,624	0,630	**Нет**	0,376	1,643	0,244
18	Нет	Нет	0,953	0,166	Да	0,047	6,166	-1,445
19	Нет	Нет	0,962	0,271	Да	0,038	6,739	-1,558
20	Нет	Нет	0,663	0,507	Да	0,337	1,859	-0,326
21	Нет	Нет	0,984	0,904	Да	0,016	9,160	-1,989
22	Нет	Нет	0,976	0,545	Да	0,024	7,916	-1,776
23	Нет	Нет	0,997	2,850	Да	0,003	14,167	-2,726
24	Нет	Нет	0,979	0,643	Да	0,021	8,279	-1,839

Проверка качества дискриминантной модели. Общим подходом к проверке качества процедур классификации является переклассификация объектов обучающей выборки, для которых принадлежность к классу известна, по разработанным правилам классификации, в том числе с использованием построенной модели. Этот подход мы рассматривали применительно к логистической регрессии (см. табл. 1.20). Он же применяется в дискриминантном анализе (табл. 3.8).

Кроме того, для проверки качества дискриминантной модели используются статистические критерии. Наиболее распространенным критерием является статистика λ-Уилкса (лямбда), измеряющая качество разделения классов: чем ближе значение λ к нулю, тем различение выше. На основе λ-Уилкса вычисляются критерий X^2 (хи-квадрат) и соответствующий ему уровень значимости.

Пример 3.2 (продолжение)

Таблица 3.8

Результаты переклассификации обучающей выборки

Актуальная группа	Предсказанная группа		
	нет	да	всего
Нет	10	2	12
Да	1	11	12
Нет (%)	83,3	16,7	100,0
Да (%)	8,3	91,7	100,0

Всего классифицировано правильно 87,5 % обучающей выборки.

Последовательный отбор дискриминантных переменных может производиться аналогично тому, как отбираются независимые переменные при построении уравнения множественной линейной регрессии (см. с. 36–37). Основным критерием при выборе переменных для включения в уравнение канонических дискриминантных функций является уменьшение значения статистики λ-Уилкса, а также ряд критериев, помогающих определить информационную ценность дискриминантных переменных[1].

Самостоятельная работа

Классификация сенаторов США по их внешнеполитическим взглядам проводилась американским политологом Б. А. Бардес[2] на основе результатов их голосования по вопросам помощи иностранным государствам. Переменные, фиксирующие результаты голосования, имеют три значения: 3 – «за»; 1– «против», 2 – «воздержался» или отсутствовал при голосовании. Все голосования разделены на 6 групп; для каждой группы вычислена дискриминантная переменная как среднее значение голосований по соответствующим поводам каждого сенатора.

Дискриминантные переменные:

CUTAID – за сокращение фондов помощи;

RESTRICT – за добавление ограничений в программу помощи;

CUTASIA – за сокращение помощи азиатским государствам;

MIXED – за помощь некоторым странам, но никакой помощи коммунистам;

ANTIYUGO – за неоказание помощи Югославии;

ANTINEUT – за неоказание помощи нейтральным странам.

Коэффициенты нестандартизированных канонических переменных приведены в табл. 3.9; выделенные классы и характеристики обучающей выборки – в табл. 3.10; координаты центроидов в пространстве нестандартизированных канонических дискриминантных функций – в табл. 3.11.

---

[1] *Клекка У. Р.* Указ. соч. С. 122–130.

[2] Там же. С. 84–86.

Таблица 3.9

Коэффициенты нестандартизированных канонических дискриминантных функций и классификационных функций

Переменная	Нестандартизованные функции			Классификационные функции			
	f_1	f_2	f_3	h_1	h_2	h_3	h_4
CUTAID	0,09	-0,52	1,62	13,04	6,28	9,06	11,94
RESTRICT	0,79	-1,12	-0,33	5,76	1,60	1,49	2,42
CUTASIA	-4,60	-1,12	-1,14	20,06	59,29	33,45	15,89
MIXED	-0,70	-1,32	1,14	37,02	42,91	36,76	33,25
ANTIYUGO	-1,11	1,11	0,38	-2,64	7,48	2,63	0,65
ANTINEUT	1,43	1,04	0,20	8,56	-3,52	5,54	11,90
Constant	5,42	3,57	-4,38	-77,59	-146,88	-87,33	-69,19

Таблица 3.10

Выделенные классы сенаторов и обучающая выборка

Класс	Количество сенаторов	Стратегия голосования
*GR*1	9	В целом за помощь иностранным государствам
*GR*2	2	В целом против помощи иностранным государствам
*GR*3	5	Против помощи неплатежеспособным государствам
*GR*4	3	Антикоммунисты

Таблица 3.11

Центры классов обучающей выборки в пространстве дискриминантных функций

Переменная	*GR*1	*GR*2	*GR*3	*GR*4
CUTAID	1,42	3,00	2,20	2,10
RESTRICT	1,94	1,00	2,00	2,33
CUTASIA	1,00	3,00	2,00	1,33
MIXED	2,67	2,00	1,80	1,67
ANTIYUGO	1,56	2,50	2,60	3,01
ANTINEUT	1,26	1,67	2,13	2,44

Задание

Классифицируйте сенатора *A*, который по результатам голосований получил следующие значения дискриминантных функций:

CUTAID = 2; RESTRICT = 3; CUTASIA = 2; MIXED = 3; ANTIYUGO = 3; ANTINEUT = 2.

Литература

Бессокирная, Г. П. Дискриминантный анализ для отбора информативных переменных / Г. П. Бессокирная // Социология : 4М. 2003. № 16.

Большев, Е. С. Дискриминантный анализ в прогнозировании поведения неопределившихся избирателей / Е. С. Большев // Социология : 4М. 2009. № 29.

Бююль, А. SPSS: Искусство обработки информации / А. Бююль, П. Цёфель. М., 2002.

Елисеева, И. И. Группировка, корреляция, распознавание образов / И. И. Елисеева, В. О. Рукавишников. М., 1977.

Елисеева, И. И. Основные процедуры многомерного статистического анализа / И. И. Елисеева, Е. В. Семенова. СПб., 1993.

Клекка, У. Р. Дискриминантный анализ / У. Р. Клекка // Факторный, дискриминантный и кластерный анализ / под ред. И. С. Енюкова. М., 1989.

Климова, С. Г. Анализ настроений методом структурно-логической типизации / С. Г. Климова, Е. Г. Галицкая // Социология : 4М. 2010. № 30.

Крыштановский, А. О. Анализ социологических данных / А. О. Крыштановский. М., 2007.

Наследов, А. Д. SPSS: Компьютерный анализ данных в психологии и социальных науках / А. Д. Наследов. СПб., 2007.

Олдендерфер, М. С. Кластерный анализ / М. С. Олдендерфер, Р. К. Блэшфилд // Факторный, дискриминантный и кластерный анализ. М., 1989.

Орлов, А. И. Заметки по теории классификации / А. И. Орлов // Социология : 4М. 1991. № 2.

Сатаров, Г. А. Расчет рейтингов законодателей: (Консерватизм и радикализм на II Съезде народных депутатов СССР) / Г. А. Сатаров, С. Б. Станкевич // Демократические институты в СССР: проблемы и методы исследования / сост. И. В. Задорин ; науч. ред. О. М. Маслова. М., 1991.

Сивуха, С. В. Социальные детерминанты самооценки успеха / С. В. Сивуха, М. Титма // Социальное расслоение возрастной когорты. М., 1997.

Татарова, Г. Г. Основы типологического анализа в социологических исследованиях / Г. Г. Татарова. М., 2004.

Татарова, Г. Г. Типологический анализ для реконструкции социальных типов работников / Г. Г. Татарова, Г. П. Бессокирная // Социол. исслед. 2011. № 7.

Терещенко, О. В. Классификация социологических объектов методами многомерного статистического анализа // Вестн. БГУ. Сер. 3. 1994. № 2.

Типология и классификация в социологических исследованиях / отв. ред. В. Г. Андреенков, Ю. Н. Толстова. М., 1982.

Черныш, М. Ф. Опыт применения кластерного анализа / М. Ф. Черныш // Социология : 4М. 2000. № 12.

Глава 4

ОСНОВЫ АНАЛИЗА СОЦИАЛЬНЫХ СЕТЕЙ

Анализ социальных сетей (АСС), или сетевой анализ, – это группа методов анализа связей между социальными акторами, представляемых в виде различных отношений, взаимодействий и потоков ресурсов, другими словами, – методов анализа реляционных данных. Сетевой анализ – одно из наиболее перспективных и многообещающих направлений междисциплинарных исследований, которое используется для изучения самых разных социальных явлений, таких как учет социальных связей при поиске работы; роль сетей в экономической жизни домохозяйств и миграционных процессах; организационные и межорганизационные структуры взаимодействия; власть и влияние; социальный капитал и многое другое.

Аналитический аппарат АСС уникален по своим возможностям изучения и моделирования не атрибутивных, а реляционных (структурных) данных, которые отражают свойства системы социальных отношений и не сводятся к характеристикам акторов. Основу сетевого анализа, восходящего к социометрическим исследованиям Я. Л. Морено[1], составляют математическая теория графов, матричная алгебра и теория вероятностей.

Категориальный аппарат сетевого анализа включает в себя как социологические термины, так и их формальные математические эквиваленты. В специализированной литературе одинаково часто можно встретить и те и другие термины. Рассмотрим коротко концептуальные понятия сетевого анализа и их общепринятую интерпретацию.

Актор – субъект социального действия, который в зависимости от задачи может быть как отдельным индивидом, так и коллективным со-

[1] *Морено Я. Л.* Социометрия: Экспериментальный метод и наука об обществе. М., 2004.

циальным объединением любого масштаба и размера – от студенческой группы до международного объединения различных организаций и целых стран. Графически и математически актор представляется как *вершина* или *узел* графа[1]. Вершина графа может обозначаться буквой (например, A), номером (например, i или n_i – от англ. *node*), количество вершин в графе – буквой N.

Связь – взаимодействие или отношения между двумя акторами. Содержание (тип взаимодействий) и форма (измерение) связи зависят от целей и задач исследования.

Основные типы социальных взаимодействий, которые являются предметом изучения в исследованиях социальных сетей, следующие:

- аффективное оценивание (выражение отношений дружбы, любви, доверия);
- обмен материальными ресурсами (бизнес-трансакции, денежные займы, ссуды) и нематериальными ресурсами (коммуникация, получение/передача информации);
- связи идентификационного характера, выражающие чувство принадлежности человека к социальной группе;
- перемещение физическое (с места на место) и социальное (между позициями и статусами);
- формальные, административные связи (власть и влияние);
- отношения родства (брак, происхождение).

В зависимости от способа измерения связи между акторами могут быть:

- ненаправленными, дихотомическими (предполагают фиксацию наличия (1) или отсутствия (0) связи между акторами). Ненаправленная связь между двумя вершинами графа изображается линией и называется *ребром* (*edge*) (рис. 4.1);
- направленными (предполагают фиксацию направления связи между акторами: ($A \to B$) или (A, B)). Направленная связь от одной вершины к другой в терминах теории графов называется *дугой* (*arc*), на графе изображается линией со стрелкой (рис. 4.2);
- маркированными, или означенными (предполагают фиксацию положительных и отрицательных связей акторов). Характер связи обозначается на графе знаками «+» и «–» соответственно (рис. 4.3);
- оценочными, или взвешенными (предполагают фиксацию интенсивности связи между акторами, например, через частоту контактов, денежный торговый оборот между странами и т. д.). Интенсивность связи указывается на графе числом (рис. 4.4).

[1] В отличие от анализа статистических связей, в анализе социальных сетей вершины графа соответствуют «объектам» – членам сети, а не переменным.

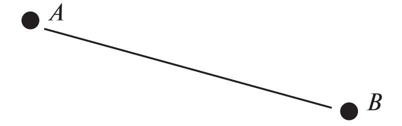


Рис. 4.1. Ребро графа, связывающее вершины *A* и *B*

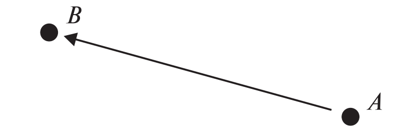


Рис. 4.2. Дуга графа, связывающая вершины *A* и *B*

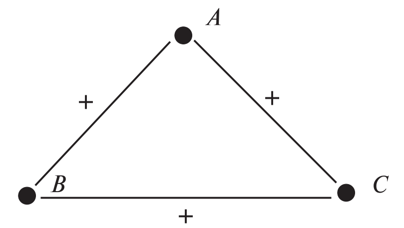


Рис. 4.3. Положительные связи между акторами *A*, *B* и *C*

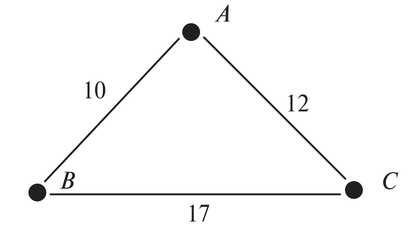


Рис. 4.4. Интенсивность связей между акторами *A*, *B* и *C*

Таким образом, с помощью ребер и дуг схематически изображают пути, по которым могут протекать различные виды социальной активности акторов сети. Ребра и дуги обобщенно будем называть линиями (*line*) и обозначать *l*; общее количество линий на графе – *L*.

Диада – это простейшее социальное объединение, состоящее из двух акторов и связей между ними. На математическом языке диада – это подграф из двух вершин *A* и *B* с заданными на нем ребрами и/или дугами. Отношения в диаде представляют собой множество упорядоченных пар, где на первом месте стоит вершина, из которой выходит данная дуга, а на втором месте – номер вершины, в которую она входит. Диады могут обозначаться (*A*, *B*) или $A \to B$, причем первое обозначение может использоваться для любых связей, а второе – только для направленных.

Выделяют три изоморфных класса диад:

M (*mutual*) – взаимная диада, в которой одновременно представлены два отношения $A \to B$ и $B \to A$ (см. рис. 4.1);

A (*asymmetric*)– асимметричная диада, в которой имеет место только одно из отношений $A \to B$ или $B \to A$ (см. рис. 4.2);

N (*nought*)– нулевая диада, в которой связей между вершинами *A* и *B* нет.

Для анализа социальных сетей диада имеет фундаментальное значение, так как именно на основе парных выборов формируются графы и матрицы, репрезентирующие паттерны социальных отношений в сети. Диада является также основной единицей статистического анализа социальных сетей.

Триада – социальное объединение, состоящее из трех акторов и возможных связей между ними (подграф из трех вершин). Особенно интересными для анализа, с точки зрения содержательной интерпретации, является класс *транзитивных триад* (от англ. *transition* – переход). Транзитивными являются триады, в которых если актор *i* выбирает актора *j* и актор *j* выбирает актора *k*, то актор *i* также выбирает актора *k*. Другими словами, когда «друг моего друга – мой друг» и «враг моего друга – мой враг». Транзитивность объясняет формирование связей в триаде, позволяя понять такие групповые эффекты, как сплоченность и поляризация, и является универсальным механизмом образования социальных связей.

Отношения в триаде являются сбалансированными, если акторы *A* и *B* выбирают друг друга и одновременно выбирают или не выбирают актора *C*. Если *A* и *B* различаются в выборе *C*, отношения в триаде несбалансированные. На инструментальном уровне различать сбалансированные и несбалансированные триады можно по количеству отрицательных связей: в сбалансированной триаде их нет или четное число (две), в несбалансированной – нечетное (одна).

Подгруппа (в терминах теории графов – *подграф*) – это обособленная часть целостной сетевой структуры. Подгруппа состоит из некоторого подмножества акторов социальной сети и связей между ними. Примерами подгрупп являются диада, триада, клика, клан, плекс и др. Их изучением занимается особое направление анализа социальных сетей – анализ связанных подгрупп.

Социальная сеть – это множество социальных акторов и определенное на нем множество социальных отношений между ними. Социальная сеть может быть представлена в виде графа и матрицы (*социоматрицы*).

Различают два основных типа социальных сетей:

- личная, или эгоцентрическая, сеть;
- полная, или социоцентрическая, сеть.

Эгоцентрическая социальная сеть (эго-сеть) ограничена личными связями (родственными, профессиональными, дружескими, соседскими и т. п.) одного актора (эго). Эго-сеть называют также частной, чтобы подчеркнуть отличие от полной сети, т. е. сети отношений между всеми членами группы.

Социоцентрическая сеть описывается полной структурой ролевых отношений в некотором сообществе, часто ограниченном формальной рамкой. Стандартная и наиболее часто используемая модель социальной сети предполагает идентичность характеристик акторов. В этом случае каждая вершина может быть связана входящими и исходящими ребрами с любыми другими вершинами данного графа. Например, вершинами могут быть работники одной организации, а ребра будут соответствовать дружеским отношениям между ними.

С точки зрения теорий обмена и социального капитала социальная сеть – это совокупность не столько акторов, сколько позиций, которые они занимают, а также связей и ресурсов, циркулирующих между данными позициями. Таким образом, основные компоненты сети могут быть структурными и ресурсными. *Структурный* компонент подразумевает конфигурацию акторов и обусловленные этой конфигурацией связи между ними. *Ресурсный* компонент – это вид обмениваемых ресурсов и дифференциация между позициями на основе такого обмена.

Соответственно компонентам сети в АСС различают два основных вида переменных: структурные и атрибутивные. Ключевую роль в анализе социальных сетей играют структурные переменные, которые фиксируют или измеряют связи и контакты индивидов, включая обмен ресурсами, а также связи и контакты между группами акторов. Атрибутивные переменные, измеряющие индивидуальные свойства акторов (пол, умственные способности, располагаемые ресурсы, размер организации и др.), также используются в сетевом анализе, но выполняют они скорее вспомогательную роль, облегчая интерпретацию полученных результатов.

Как отмечает Б. Уэлман, сетевой анализ не приемлет трактовку социального поведения только как результата индивидуальных характеристик или атрибутов акторов, дополняя их включенностью в структуру социальных отношений[1]. Р. Берт указывает, что сетевая теория строит свое объяснение на паттернах отношений; она фиксирует причинные факторы в основных принципах построения социальной структуры общества, минуя «иллюзорно-значимые атрибуты людского бытия»[2].

Аналитические принципы сетевого анализа. Основные аналитические принципы, касающиеся наиболее важных механизмов образования и функционирования паттернов социальных отношений современного общества, выделили Б. Уэлман и С. Берковитц в работе «Социальные структуры: Сетевой подход»[3]. Понимание, учет этих механизмов существенны для эффективного объяснения социальных, экономических, политических и других процессов современного общества.

1. Социальные связи, представляющие собой потоки ресурсов, в большинстве случаев асимметричны. Социальные каналы взаимодействия в обществе выполняют важную функцию перераспределения самых различных видов ресурсов – материальных (деньги, услуги, товары) и нематериальных (эмоции, информация). Результаты многочисленных исследований пока-

[1] *Wellman B.* Network Analysis: Some Basic Principles // Sociological Theory / R. Collins (ed.). NY, 1983. P. 165.

[2] *Burt R.* Toward a Structural Theory of Action: Network Models of Social Structure, Perception, and Action. NY, 1982. P. 106.

[3] *Berkowitz S. D., Wellman B.* Social Structures: A Network Approach. Cambridge, 1988. P. 19–46.

зывают, что преобладающее большинство связей, типа патрон – клиент, дети – родители, муж – жена, учитель – ученики и другие, предполагают однонаправленную поставку различного рода ресурсов между акторами.

2. Структура распределения ресурсов в сети определяет позиции акторов в ней. Сетевой подход позволяет исследователям идентифицировать властные отношения между социальными акторами и моделировать связи между поведением актора и его структурной позицией. В терминах сетевого анализа власть (как противоположность зависимости) производна от степени контроля за важными ресурсами: контроль одного актора означает, что другой актор имеет мало альтернативных источников получения ресурса. Другими словами, один актор контролирует или опосредует доступ другого актора к ресурсам или благам. Например, работодатели могут контролировать определенные ресурсы, приобретать их посредством процессов обмена и добиваться желаемых благ, тем самым повышая зависимость наемных работников от себя. Акторы, опосредующие доступ к дефицитным ресурсам, являются *посредниками*. Значимость посредника, «сила» его структурной позиции зависит от того, какую роль он выполняет в данной социальной структуре. Р. Берт следующим образом классифицирует функции посредников[1]:

- *координатор* – обеспечивает передачу ресурса между членами своей группы;
- *консультант* – обеспечивает передачу ресурса для группы, членом которой он не является;
- *представитель* – обеспечивает передачу ресурса из своей группы в другую;
- *связной* – обеспечивает передачу ресурса из одной группы в другую, не являясь членом ни одной из них.

3. Универсальные механизмы образования связей (сходство, близость, транзитивность) предопределяют образование сплоченных подгрупп в сети. Общество не строится на основе случайных связей между людьми, но имеет тенденцию расчленяться на ряд сплоченных групп, формирование которых во многом предопределяют три универсальных механизма образования связей. Первый механизм – это *сходство характеристик* (*гомофилия*). Теории гомофилии в целом сводятся к тому, что сходство социально-демографических характеристик акторов (пол, возраст, профессия, место жительства и т. д.) в значительной степени облегчает коммуникацию и установление доверия между ними, т. к. их действия и поведение понятны и предсказуемы. Второй механизм – *территориальная и / или электронная близость* (*соседство*). Под электронной близостью имеется в виду наличие и доступность средств электронной связи (компьютер, мобильный теле-

[1] *Burt* R. Brokerage and Closure: An Introduction to Social Capital. Oxford, 2004. P. 95.

фон, доступ в Интернет и др.). Близость (соседство) значительно увеличивает вероятность встреч и, соответственно, взаимодействия индивидов друг с другом. Третий универсальный механизм образования связей, который необходимо учитывать при анализе взаимодействий, – это *тенденция образования транзитивных связей в триаде*. По меткому замечанию С. Милграма, без данного свойства весь мир представлял бы собой один гигантский кластер[1].

4. Сплоченные подгруппы в сети обычно накладываются друг на друга (перекрываются). В определенный момент времени каждый социальный актор обладает конечным количеством связей, что обусловливается его физическими, эмоциональными, материальными и другими возможностями. Данные связи по своей природе множественны, т. е. предполагают различные отношения между акторами (дружба, родство, знакомство, членство в организации и др.). В результате социальные акторы входят в различные социальные объединения, сплоченные подгруппы, одновременно связывая их между собой, другими словами, обеспечивая их пересечение. С точки зрения сетевого анализа количество и качество (плотность, связанность и т. д.) сплоченных подгрупп в сети является одной из важнейших характеристик социальной структуры в целом, позволяющей объяснять такие сложные процессы, как распространение инноваций, инфекционных заболеваний и т. д.[2]

5. Доступ к дефицитным ресурсам, разрешение кризисных ситуаций, мобилизация коллективных действий эффективнее всего достигаются посредством неформальных связей социального актора. В таких ситуациях неформальные связи воплощают в себе стратегию кооперации в интересах уменьшения неопределенности и усиления безопасности и стабильности. Другими словами, если обладание ресурсами в стабильном и предсказуемом обществе может позволить людям быть независимыми друг от друга, то противоположная ситуация заставляет искать личные связи с теми, кто может обеспечить поддержку и защиту[3]. Например, неформальные связи особенно важны в условиях дефицита финансовых ресурсов и отсутствия материальных ресурсов, которые могли бы использоваться в качестве залога. В такой ситуации именно личные знакомства служат залогом под финансовые кредиты[4].

---

[1] *Милграм С.* Эксперимент в социальной психологии. СПб., 2000. С. 125.

[2] *Kadushin C.* Introduction to Social Network Theory [Electronic resource]. M., 2004. URL: http://www.communityanalytics.com/Portals/0/Resource_Library/Social%20Network%20тTheory_Kadushin.pdf.

[3] *Динелло Н.* От плана к клану: социальные сети и гражданское общество [Электронный ресурс].URL: http://www.prof.msu.ru/publ/book3/din.htm.

[4] *Molina J. L.* The Informal Organizational Chart in Organizations: An Approach from the Social Network Analysis // Connections: Electronic journal. 2001. № 24(1) P. 78–91. URL: http://www.insna.org/PDF/Connections/v24/2001_I-1_78-91.pdf.

Основная идея статьи М. Грановеттера «Сила слабых связей», принесшей мировую известность своему автору, состоит в том, что сети выполняют исключительно важную роль установления соответствия между спросом и предложением на рынке труда, т. к. в ходе межличностной коммуникации передается информация, не циркулирующая по публичным каналам. Более того, работа, найденная через сети, как правило, более высокого качества в смысле содержания, оплаты и условий труда[1].

Таким образом, сетевой анализ объясняет разнообразные аспекты человеческого поведения через социальные связи, отрицая тем самым, по крайней мере, четыре популярные исследовательские стратегии: 1) редукционистское объяснение поведения личными качествами индивидов; 2) объяснение, основанное на ценностях, идеях, когнитивных картах и т. п. (в этом отношении структурализм следует отличать от структурного анализа социальных сетей); 3) технологический и материальный детерминизм; 4) объяснение, основанное на причинных регрессионных моделях.

Определение границ социальной сети. Процесс реконструкции и анализа социальной сети начинается с определения ее границ, т. к. при осуществлении сбора информации в сетевых исследованиях довольно часто возникает проблема выбора объектов, включаемых в сеть. Очевидно, что упущение необходимых элементов или произвольное очерчивание границ могут ввести исследователя в заблуждение и профанировать результаты. Принято выделять два основных подхода к определению границ социальной сети: реализм и номинализм.

Реализм основан на субъективном восприятии акторами границ сети и своего членства в ней. Данная стратегия получила название «актороцентрической» и заключается в сборе данных обо всех (либо о каких-то конкретных) взаимодействиях, в которые включен определенный актор. Эта стратегия особенно часто используется при реконструкции социальной сети по результатам опроса респондентов. Показательно в этом отношении исследование личных сетей граждан США, которое осуществил американский социолог К. Маккарти[2]. Он обращался к своим респондентам с просьбой сначала назвать 60 членов своей личной сети, а затем определить, кем они приходятся респонденту, и оценить связи между ними. Цель данного исследования была в том, чтобы изучить типы, виды и количество связей, в которые включен каждый конкретный актор, не навязывая своей точки зрения, т. е. максимально устранив влияние исследователя.

[1] *Granovetter M. S.* The Strength of Weak Ties // American Journal of Sociology. 1973. № 78. P. 1367.

[2] *MacCarty C.* Structure in Personal Networks // Journal of Social Structure [Electronic resource]. 2002. Vol. 3, № 1. URL: http://zeeb.library.cmu.edu:7850/JoSS/McCarty/McCarty.

Реалистический подход достаточно часто применяется в исследованиях социальных совокупностей, границы которых размыты или акторы которых труднодоступны для исследования (бизнес-сообщества, преступные группировки, наркоманы, коллекционеры и т. д.). В этом случае в качестве основной выборочной техники используется метод «снежного кома», вследствие чего его стали называть методом *сетевой выборки*.

Номинализм – исследовательская традиция, базирующаяся на теоретических представлениях исследователя. В этом случае сеть выделяется исследователем самостоятельно по некоторому критерию, исходя из целей исследования, и сама по себе «не имеет онтологически независимого статуса»[1]. Если реалистическая стратегия заключается в сборе данных обо *всех взаимодействиях*, в которые включен определенный актор, то номиналистическая стратегия направлена главным образом на получение данных о взаимодействиях *всех акторов* сети. Зачастую номиналистический подход оказывается единственно возможным при построении социоцентрической сети, если сбор данных осуществляется методами анализа документов, наблюдения, эксперимента.

Номиналистический подход предполагает, как правило, сплошное обследование акторов сети и их взаимодействий между собой. Выборочный отбор в рамках данного подхода обычно не производится, т. к. единицы анализа (акторы и их связи) не являются независимыми. Поэтому особенно эффективным данный подход является в случае исследования небольшого, замкнутого сообщества (школьный класс, клуб, организация и др.), где включение в анализ всех акторов сети не представляет особой сложности. В качестве критериев включения в сеть могут служить участие в общественной организации, членство в профессиональном сообществе, принадлежность к элите и др. Например, в исследовании распространения информации о новых лекарствах в локальном сообществе лечащих врачей в выборку могут быть включены все местные врачи либо врачи определенной специализации.

Таким образом, в рамках сетевого анализа не существует жестких схем и установок относительно реконструкции социальной сети. Способы определения границ сети зависят, во-первых, от характеристик изучаемой структуры и ее элементов, во-вторых, от того, какие именно социальные связи необходимо включить в анализ, в-третьих, от того, каким методом мы собираемся получить информацию о сетевых взаимодействиях.

Сбор данных в сетевых исследованиях. Выбор метода сбора данных зависит, в первую очередь, от свойств и характеристик социальных субъектов, между которыми необходимо установить наличие взаимодействий, а также от материальных и технических ресурсов исследователя. Наибо-

[1] *Laumann E. O., Marsden P. V., Prensky D.* The Boundary Specification Problem in Network Analysis // Research Methods in Social Network Analysis / L. C. Freeman, D. R. White, A. K. Romney (eds.). New Brunswick ; New Jersey, 1992. P. 65.

лее популярными методами в исследованиях социальных сетей являются опрос и анализ документов.

Опрос. Специфика опроса в исследовании межличностных связей определяется необходимостью получения двух видов данных – атрибутивных и реляционных (структурных). Бланк анкеты или интервью включает в себя, как правило, два блока вопросов: в первом фиксируются социально-демографические и другие значимые характеристики респондента, во втором осуществляется реконструкция связей респондента с акторами изучаемой сети. В процессе реконструкции связей респондента его обычно просят указать имена и некоторые атрибутивные характеристики людей, с которыми он взаимодействует. Данный способ получения информации от респондентов называется «генератор имен». Наиболее активно этот метод используется при изучении разных форм социального обмена: интенсивности общения, обмена личными проблемами, займа денег и др.

Список акторов, который респондент создает в процессе опроса, называется списком свободного припоминания или свободного выбора, если количество названных имен не ограничено. Если респондентов просят назвать фиксированное количество людей, которых они считают, например, своими лучшими друзьями, получают список фиксированного выбора. Использование списков фиксированного выбора резко снижает надежность полученных данных, однако улучшает их статистические свойства.

В исследованиях индивидуального социального капитала довольно активно используются и другие методы, например «генератор позиций» и «генератор ресурсов». «Генератор позиций» основан на измерении профессионального положения участников эго-сети актора. В ходе опроса респондентов просят указать, есть ли среди их знакомых врачи, юристы, строители, предприниматели и др. «Генератор ресурсов» соответственно основан на измерении наличия дефицитных ресурсов (деньги, жилье, знание компьютерных программ, законов и т. д.) участников эго-сети актора[1].

Анализ документов. Использование опросных методов в исследовании социальных сетей влечет за собой ряд концептуальных проблем, связанных с субъективными факторами. Изучение документальных источников – архивных записей, исторических и литературных произведений, статистических и отчетных данных, журнальных и газетных статей, следственных материалов и др. – позволяет избежать этих проблем благодаря тому, что в сеть преобразуется текстовая информация. Документальный подход основан на предположении, что, упоминая имена персонажей (акторов) и события, авторы текста в некотором смысле связывают себя с ними[2], считают их значимыми в положительном или отрицательном кон-

---

[1] *Wasserman S., Faust K.* Social Network Analysis: Methods and Applications. Cambridge, 1994. P. 86.

[2] См., напр.: *Батыгин Г. С., Градосельская Г. В.* Сетевые взаимосвязи в профессиональном сообществе социологов: методика контент-аналитического исследования // Социол. журн. 2001. № 1.

текстах. Задача, следовательно, заключается в регистрации такого рода упоминаний.

С точки зрения валидности и надежности основным недостатком анализа документальных источников является естественная фрагментарность получаемых данных. Данную проблему в некоторых случаях удается разрешить, используя метод триангуляции. Например, при исследовании взаимодействия в научном сообществе одновременно используются связи цитирования в научной печати, принадлежность к определенным организациям, посещение конференций, коммуникации по электронной почте[1].

Другие методы сбора данных. При реконструкции социальных сетей, кроме опроса и анализа документов, используются также другие методы. Для изучения процессуального характера социальной жизни в относительно небольших социальных сетях, где осуществляется взаимодействие лицом к лицу, может использоваться *наблюдение*. В исследованиях эго-сетей применяется метод *анализа личных дневников* респондентов. *Метод социального эксперимента* чаще может использоваться при изучении небольших групп и в исследованиях «тесного мира»[2]. В последнее время значительно упростился процесс технической регистрации взаимодействий сетевых акторов, использующих электронные средства связи для взаимодействий друг с другом, в результате чего все больше исследований организационной, межорганизационной и научной коммуникации проводятся на основе *анализа потоков электронной почты, посещения гиперссылок* и т. п. С точки зрения сетевого подхода гиперссылка является не просто технологическим инструментом, но также новым коммуникационным каналом, посредством которого веб-сайты в сети связаны друг с другом (в данном случае актором сети выступает веб-сайт, принадлежащий индивиду или организации).

Описательные методы анализа социальных сетей. Богатый математический аппарат сетевого анализа, накопленный за многолетнюю историю его становления и развития, позволяет реконструировать разнообразные модели социальных взаимодействий – от простых до самых сложных. Различают две основные группы методов сетевого анализа: 1) методы описательного анализа, разрабатываемые в рамках теории графов и матричной алгебры; 2) методы стохастического анализа, разрабатываемые преимущественно в рамках теории вероятностей и теории статистического вывода.

Представление данных в сетевом анализе. Описательный сетевой анализ позволяет визуализировать изучаемую сетевую структуру, формализовать положение индивида в сети; изучить сплоченные подгруппы и их основные характеристики, определить сбалансированность триад в сети

[1] *Friedkin N. E.* A Structural Theory of Social Influence. London, 1998. P. 114.
[2] *Милграм С.* Указ. соч.

и др. В силу относительной простоты расчетов и интерпретации данные методы наиболее часто используются в исследованиях социальных сетей.

Подчеркнем еще раз значимость в сетевом анализе методов визуализации данных, значительно облегчающих восприятие социальных сетей и позволяющих исследователю не только по-новому открыть для себя социальную структуру, но и передать это знание другим.

В анализе социальных сетей структурные переменные представляются в виде социоматрицы – квадратной таблицы, строки и столбцы которой соответствуют одним и тем же акторам (табл. 4.1). На пересечении строки i и столбца j фиксируются числовые значения $x_{i,j}$, характеризующие связи между акторами i и j. На главной диагонали матрицы ($i = j$) значения не указываются, т. к. связь объекта с самим собой не имеет смысла. Если мы фиксируем ненаправленные связи $x_{i,j} = x_{j,i}$, матрица является симметричной, если же связи между акторами являются направленными, матрица является асимметричной, например: $x_{i,j} = 1$, $x_{j,i} = 0$. Обработка и анализ сетевых данных представляют собой преобразование матрицы с целью получить компактное описание как всей структуры взаимоотношений между членами группы, так и некоторых характеристик для каждого актора сети в отдельности.

Пример 4.1. Сеть социальных связей работников организации

Допустим, изучается специфика неформальных связей в некой организации. Работников этой организации просят указать, кому из своих коллег они прежде всего передают информацию, связанную с профессиональной деятельностью. Полученные результаты заносят в асимметричную матрицу данных (табл. 4.1): 1– передает информацию, 0 – не передает.

Таблица 4.1

Передача профессиональной информации сотрудниками организации (асимметричная матрица)

Актор	*А*	*Б*	*В*	*Г*	*Д*	*Е*	*Ж*	*З*	*И*	*К*
А		1	0	0	1	0	1	0	1	0
Б	1		1	1	1	0	1	1	1	0
В	0	1		1	1	1	1	0	0	1
Г	1	1	0		1	0	1	0	0	0
Д	1	1	0	1		0	1	1	1	1
Е	0	0	1	1	1		1	0	1	0
Ж	0	1	1	1	1	0		0	0	0
З	1	1	0	1	1	0	1		1	0
И	0	1	0	0	1	0	1	0		0
К	1	1	1	0	1	0	1	0	0	

Согласно полученным данным, актор *А* чаще всего передает информацию акторам *Б*, *Д*, *Ж*, *И* (данные, расположенные по строке). Тогда как актору *А* чаще всего передают информацию акторы *Б*, *Г*, *Д*, *З* и *К* (данные, расположенные по столбцу).

Каждой социоматрице связей взаимно однозначно соответствует граф (*G*), являющийся удобным и эффективным средством визуализации социальных взаимодействий. Он включает в себя два уровня информации: некоторое множество вершин сети $N = \{n_1, n_2 \ldots n_N\}$ и некоторое количество линий $L = \{l_1, l_2 \ldots l_L\}$, представляющих собой связи-отношения между вершинами графа. Заметим, что буквы *N* и *L* используются не только для обозначения множеств вершин и линий, но и их объема. Понятие размера сети обычно связывают не с количеством вершин *N*, а с количеством ребер и дуг *L*. Для построения графов используется специальное программное обеспечение[1].

Как отмечалось выше, связи в графе могут быть ненаправленными (ребра) и направленными (дуги). Ребра могут отражать такие отношения, как, например,«родственные связи» или совместная работа; дуги – упоминание кого-то в качестве друга, оказание помощи, предоставление информации (как в нашем примере) и т. д. Граф с заданными на нем дугами называется *ориентированным* или *орграфом*.

Пример 4.1 (продолжение)

Полученной матрице связей (см. табл. 4.1) взаимно однозначно соответствует орграф связей, представленный на рис. 4.5.

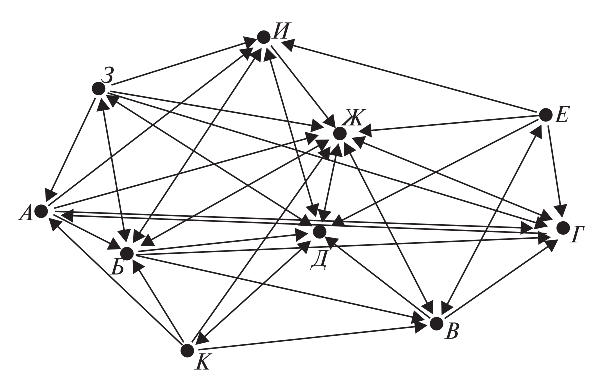


Рис. 4.5. Орграф передачи профессиональной информации работниками организации

[1] Например, программа *Ucinet*, авторы которой С. Боргатти, М. Эверетт и Л. Фриман, содержит целый набор алгоритмов визуализации структурных данных (многомерное шкалирование, анализ соответствий и др.) и позволяет редактировать полученные визуальные структуры. Данная программа, являющаяся своего рода стандартом в области компьютерной обработки структурных данных, находится в свободном доступе на сайте *www.analytictech.com/ucinet*.

***Плотность сети* (*dencity*)** – показатель того, насколько интенсивно акторы связаны в сети между собой. Вычисление данного показателя начинается с установления размера сети – подсчета количества реально присутствующих и максимально возможных ребер или дуг в зависимости от типа изучаемых связей.

Максимально возможное количество ребер в любом неориентированном графе равно $N = (N - 1)/ 2$, где N – количество вершин графа, т. к. если актор A связан с актором B симметричной связью, то и B связан с A. Если матрица связей является асимметричной (а граф, соответственно, ориентированным), то из того, что A связан с B, не обязательно следует, что B связан с A и максимально возможное количество дуг в орграфе равно $N(N - 1)$.

Плотность неориентированного графа равна доле реализованных ребер, т. е. отношению числа наличных ребер L к числу потенциально возможных: $\Delta = 2L/N(N-1)$.

Плотность орграфа – доля реализованных дуг, т. е. отношение числа наличных дуг к числу потенциально возможных: $\Delta = L/N(N-1)$.

Показатель плотности сети может принимать значения из интервала [0, 1] – чем больше связей в сети отношений присутствует, тем ближе значение плотности к единице. Плотность сети может интерпретироваться как *скорость распространения в ней информации*, как *мера социальных возможностей или ограничений*. Высокая плотность эго-сети говорит об избыточности связей эго и негативно сказывается на общем объеме его социального капитала[1].

Пример 4.1 (продолжение)

Максимально возможное количество связей в рассматриваемом орграфе равно 90 ($N = 10$, следовательно, $N(N - 1) = 90$). Реально в матрице присутствует 51 дуга. Тогда плотность составляет 0,57 (51/90).

Значение плотности 0,57 свидетельствует о довольно плотной сети связей работников организации, где каждый сотрудник в среднем передает информацию пяти своим коллегам.

***Связность сети* (*connectivity*)** – показатель доступности, достижимости акторов в сети связей. Для того чтобы определить понятие связности, нам понадобится несколько новых определений.

Вершины, связанные ребром или дугой, называются *смежными*. Ребра или дуги, примыкающие к одной вершине, также называются *смежными*. В свою очередь, ребро или дуга, примыкающее к вершине, называется *инцидентным* ей. Вершина, к которой примыкает ребро или дуга, называется *инцидентным* этой вершине.

---

[1] *Hanneman R. A., Riddle M.* Introduction to Social Network Methods [Electronic resource]. URL: http://www.faculty.ucr.edu/~hanneman/nettext/.

Характеристикой вершины n_i является ее *степень* (*degree*) $s(n_i)$, которая равна количеству соединенных с ней дуг или ребер. В орграфах измеряются также исходящая $s_0(n_i)$ и входящая $s_I(n_i)$ степени вершины, равные соответственно числу дуг, исходящих из вершины и входящих в нее.

Маршрут (*walk*) – любая последовательность инцидентных вершин и ребер (дуг). Другими словами, это такая последовательность ребер (дуг), в которой конец одного ребра (дуги) является началом другого. При этом одни и те же ребра (дуги) могут встречаться в маршруте несколько раз и направлением дуг можно пренебречь. Таким образом, «путешествуя» по маршруту из одной вершины в другую, можно двигаться «против стрелок». Частными случаями маршрута являются путь (*path*) и цепь (*trail*). *Цепью* называется маршрут, в котором ребра (дуги) не повторяются; как и в маршруте, направлением дуг в цепи пренебрегают. Цепь является *простой*, если в ней не повторяются не только ребра, но и вершины. *Путь* представляет собой цепь, в которой дуги могут «проходиться» только от начала к концу, т. е. по направлению стрелки. *Простой путь* не содержит повторяющихся вершин.

Граф называется *связным*, если для любых двух вершин существует связывающий их маршрут. Актор доступен для других акторов, если в сети существуют маршруты, связывающие его со всеми другими участниками сети, независимо от того, сколько посредников они включают. Причем если данные асимметричны (граф ориентированный), возможна ситуация, когда актор *А* может достичь актора *В*, но *В* не может достичь *А*. Очевидно, что если некоторые акторы в сети не могут достичь других, то сеть является *фрагментарной*.

Для измерения степени связности графа используются показатели *достижимости акторов сети*, представляющие собой минимальное количество посредников, которых необходимо удалить, чтобы акторы не достигли друг друга. Показатели достижимости вычисляются для каждой пары акторов и сводятся в матрицу, которая содержит максимально полную информацию, характеризующую связность графа. Заметим, что для ориентированного графа матрица достижимости может оказаться асимметричной.

Пример 4.1 (продолжение)

Средняя степень вершины в изучаемом графе равна 5,1 (51/10). Орграф является связным, так как любые две вершины связаны между собой через посредников (табл. 4.2). Наиболее достижимыми, или информированными, являются акторы *Б*, *Д* и *Ж*, так как им чаще всего передают информацию (или через них другим). Об этом свидетельствуют значения от 3 до 7 вершин в соответствующих столбцах таблицы. Наименее достижимым, а значит наименее информированным, и зависимым является актор *Е*: если убрать всего одного актора *В* (см. рис. 4.5), он не получит информации совсем. Об этом свидетельствуют значения 1 в соответствующем столбце таблицы.

Таблица 4.2

Достижимость акторов сети

Актор	*А*	*Б*	*В*	*Г*	*Д*	*Е*	*Ж*	*З*	*И*	*К*
А	0	4	3	3	4	1	4	2	4	2
Б	5	0	3	5	7	1	7	2	5	2
В	4	6	0	5	6	1	6	2	4	2
Г	4	4	3	0	4	1	4	2	4	2
Д	5	7	3	5	0	1	7	2	5	2
Е	4	5	4	5	5	0	5	2	4	2
Ж	4	4	3	4	4	1	0	2	4	2
З	5	6	3	4	6	1	6	0	5	2
И	3	3	3	3	3	1	3	2	0	2
К	4	5	3	4	5	1	5	2	4	0

Орграф, в котором из каждой вершины существует путь к любой другой вершине, как в нашем примере, является *сильно связным*. В большинстве случаев высокий показатель связности сети свидетельствует о высокой плотности. Но обратное утверждение не всегда верно: довольно часто в сети с высоким показателем плотности наблюдается низкая степень связности, если в этих сетях связи существуют только внутри сплоченных подгрупп, но не между ними. Своеобразной мерой связности графа является минимальное количество вершин и ребер, при удалении которых граф окажется несвязным.

Еще одним показателем связности сети является кратчайший путь между двумя акторами – *геодезик*. Расстояние между акторами определяется длиной связывающего их маршрута, которая, в свою очередь, равна количеству входящих в него ребер (дуг). Маршрутов между двумя акторами может быть несколько, а длина кратчайшего из них и является геодезиком. В орграфе геодезик может при необходимости вычисляться с учетом направления дуг. Геодезик позволяет анализировать особенности расположения акторов в сети: чем ближе они находятся друг к другу, тем больше их возможность оказывать влияние друг на друга.

Пример 4.1 (продолжение)

В табл. 4.3 представлены геодезики для акторов организационной сети без учета направления дуг.

Таблица 4.3

Геодезики для акторов организационной сети

Актор	*А*	*Б*	*В*	*Г*	*Д*	*Е*	*Ж*	*З*	*И*	*К*
А	0	1	2	2	1	3	1	2	1	2
Б	1	0	1	1	1	2	1	1	1	2
В	2	1	0	1	1	1	1	2	2	1
Г	1	1	2	0	1	3	1	2	2	2
Д	1	1	2	1	0	3	1	1	1	1
Е	2	2	1	1	1	0	1	2	1	2
Ж	2	1	1	1	1	2	0	2	2	2
З	1	1	2	1	1	3	1	0	1	2
И	2	1	2	2	1	3	1	2	0	2
К	1	1	1	2	1	2	1	2	2	0

Максимальная длина пути равна 3, что вполне закономерно для сети с умеренным уровнем плотности 0,57. Это свидетельствует о том, что информация в данной сети распространяется довольно быстро, причем практически с «одинаковой скоростью» от любого актора.

Полезно также проанализировать информацию о том, сколько возможных путей существует от одного актора к другому. Соответствующие данные представлены в табл. 4.4.

Таблица 4.4

Количество альтернативных путей для акторов организационной сети

Актор	*А*	*Б*	*В*	*Г*	*Д*	*Е*	*Ж*	*З*	*И*	*К*
А	0	4	3	4	4	1	4	2	4	2
Б	5	0	3	6	7	1	7	2	5	2
В	5	6	0	6	6	1	6	2	5	2
Г	4	4	3	0	4	1	4	2	4	2
Д	5	7	3	6	0	1	7	2	5	2
Е	5	5	4	5	5	0	5	2	5	2
Ж	4	4	3	4	4	1	0	2	4	2
З	5	6	3	6	6	1	6	0	5	2
И	3	3	3	3	3	1	3	2	0	2
К	5	5	3	5	5	1	5	2	5	0

Большое количество альтернативных путей повышает вероятность установления контакта, в нашем случае – получения информации. В наименее выгодном положении находятся акторы *Е*, *З* и *К*.

Сила связей. Одной из важнейших характеристик социальной сети является *сила связей*, однако не существует общепринятого подхода к ее измерению. Чаще всего силу связи измеряют через частоту коммуникации, продолжительность контактов, психологическую близость, степень согласия. Сильные отношения предполагают частое общение, длительные контакты, сложные или структурно укорененные отношения, где приняты общие нормы и ценности. М. Грановеттер объединяет все эти характеристики в следующем определении: «Сила связи – это <...> комбинация времени, эмоциональной интенсивности, близости (взаимного доверия) и взаимных услуг, которые характеризуют связь»[1]. Сильные связи отличаются высокой частотой и разнообразием взаимодействий между людьми, формируют ощущение близости между ними. Поэтому они могут сопровождаться эмоциональной поддержкой и постоянным обменом информацией, имеющей отношение к общей работе или общим интересам. Слабые связи, в свою очередь, помогают не замыкаться в узком кругу, в котором циркулирует одна и та же информация. Они выводят на другие кластеры сети, прокладывая путь к новым контактам и дополнительным источникам информации[2].

Показатели центральности индивида в сети. Формализация (квантификация) положения акторов в сетевой структуре осуществляется с помощью показателей центральности. Источник идеи центральности – социодинамический закон Я. Морено, согласно которому внутри любой группы человеческие привязанности распределяются неравномерно. «Неравномерность выборов в любой группе и в каждой ситуации неустранима, – пишет Морено, – поэтому у каждой ситуации есть своя социодинамика: свои “звезды” и отверженные, у каждой группировки есть свои лидеры и лишенные взаимности»[3]. Рассмотрим наиболее распространенные подходы к анализу центральности индивида в сети.

Измерение *центральности по степени* основано на подсчете количества связей, получаемых или отдаваемых актором. Центральность по степени для входящих связей в орграфе получила интерпретацию престижа, центральность по степени для исходящих связей – влияния или экспансивности. Ф. Боначич дополнил индекс центральности по степени параметром, который он назвал «ослабляющим фактором». Этот параметр учитывает связи смежных акторов друг с другом, основываясь на положениях теории обмена Р. Эмерсона. Ф. Боначич считает, что зависимость акторов от центрального актора является тем более сильной, чем менее они связаны между собой[4].

---

[1] *Granovetter M. S.* The Strength of Weak Ties // American Journal of Sociology. 1973. № 78. P. 1367.

[2] *Granovetter M. S.* The Strength of Weak Ties: a Network Theory Revisited // Sociological Theory. 1983. № 1. P. 204.

[3] *Морено Я. Л.* Указ. соч. С. 306–307.

[4] *Bonacich P.* Power and Centrality: A Family of Measures // American Journal of Sociology. 1987. № 92(5). P. 1179.

Пример 4.1 (продолжение)

Центральность по степени для исходящих и входящих связей работников организации представлена в табл. 4.5. Показатель центральности по степени рассчитан в процентах от максимально возможного количества входящих и исходящих связей (9).

Таблица 4.5

Центральность по степени сотрудников организации

Актор	Исходящие дуги	Входящие дуги	Центральность исходящая	Центральность входящая
Д	7	9	77,78	100,00
Б	7	8	77,78	88,89
В	6	4	66,67	44,44
З	6	2	66,67	22,22
Е	5	1	55,56	11,11
К	5	2	55,56	22,22
Ж	4	9	44,44	100,00
Г	4	6	44,44	66,67
А	4	5	44,44	55,56
И	3	5	33,33	55,56

Наибольшее количество исходящих связей имеют акторы *Д* и *Б*, соответственно они имеют и наибольшее влияние в группе, выбирая, кому передать или не передать информацию. Акторы *Д*, *Б* и *Ж* имеют наибольшее количество входящих связей, что говорит об их значимости, престижности в группе, так как наибольшее количество акторов делятся с ними своей информацией.

Показатель *центральности по близости* учитывает как направленные, так и ненаправленные связи и показывает, насколько «близок» индивид к другим индивидам в сети. Если позиция центральна, то актор может максимально быстро взаимодействовать с прочими акторами. Данная позиция очень выигрышна при осуществлении коммуникации. При таком подходе центральность – это позиция, из которой необходимо делать минимальное количество шагов ко всем остальным позициям в группе. Центральность по близости привлекательна тем, что это мера *глобальной* центральности, учитывающая особенности всей сети. Основной недостаток данного вида показателей заключается в том, что они не могут быть определены для изолированных вершин в графе.

Пример 4.1 (продолжение)

Показатели центральности по близости работников организации представлены в табл. 4.6. Так как связи в нашей матрице асимметричны, количество крат-

чайших путей рассчитано и для входящих, и для исходящих путей. Центральность по близости – величина, обратная сумме расстояний от данной вершины до всех других. Как и в предыдущем случае, наиболее центральными акторами в нашей сети являются *Д* и *Ж*. Именно эти акторы быстрее всего распространят информацию, так как могут обходиться без посредников.

Таблица 4.6

Центральность по близости сотрудников организации

Актор	Сумма длин входящих кратчайших путей	Сумма длин исходящих кратчайших путей	Центральность входящая	Центральность исходящая
Д	9	12	100,00	75,00
Ж	9	14	100,00	64,29
Б	10	11	90,00	81,82
Г	12	15	75,00	60,00
А	13	15	69,23	60,00
И	13	16	69,23	56,25
В	14	12	64,29	75,00
З	16	13	56,25	69,23
К	16	13	56,25	69,23
Е	22	13	40,91	69,2

Центральность по посредничеству показывает, как часто данный актор находится на кратчайших путях (геодезиках), связывающих пары других акторов. Здесь центральность рассматривается как контроль связей между определенными позициями. Главная идея этого подхода заключается в том, что актор тем более централен, чем больше количество других акторов, между которыми он находится (чем больше маршрутов он контролирует). Индекс центральности по посредничеству интерпретируется как сумма вероятностей того, что другие акторы в своих взаимодействиях будут прибегать к посредничеству данного актора. В отличие от других индексов центральности, этот можно использовать даже в том случае, если не все вершины графа связаны друг с другом.

Пример 4.1 (продолжение)

Показатели центральности по посредничеству представлены в табл. 4.7. Центральность по посредничеству вычисляется как доля самых коротких путей, соединяющих все пары вершин, которые проходят через данную вершину. Нормированный показатель получен в результате деления исходного показателя на максимально возможное количество кратчайших путей в орграфе $(N-1)(N-2)$ и умножения на 100.

Таблица 4.7

Центральность по посредничеству сотрудников организации

Актор	Центральность по посредничеству	Нормированная центральность по посредничеству
Д	13,70	19,03
Б	11,17	15,51
В	9,95	13,82
Ж	5,78	8,03
Г	1,28	1,78
К	0,92	1,27
А	0,67	0,93
Е	0,33	0,46
И	0,20	0,28
З	0,00	0,00

В нашем примере наиболее выгодную позицию в сети по посредничеству занимают акторы *Д* и *Б*, что видно из приведенных данных. Особенно интересным здесь является то, что актор *В*, прежде не отличавшийся высокими показателями центральности, попал в тройку лидеров.

С помощью показателей центральности исследователи получают уникальную возможность изучения такого сложного и важного аспекта социальной структуры, как распределение власти и влияния. С точки зрения сетевого подхода власть не является сугубо индивидуальной характеристикой социального субъекта, она напрямую связана с его отношениями с другими людьми. Высокие или низкие показатели центральности конкретного индивида зависят от того, насколько выгодную позицию он занимает в сети отношений. Центральность по степени отражает влияние данного индивида на других, центральность по близости — скорость распространения этого влияния, а центральность по посредничеству — контролирующую, посредническую роль в этом процессе. Например, в исследовании Дж. Паджета и К. Анселла сети брачных отношений флорентийской элиты начала XV века[1] семейство Медичи было актором с наибольшей степенью центральности по посредничеству. Данное обстоятельство позволяет выявить причины, по которым представители этого семейства заняли в дальнейшем доминирующее положение во Флоренции. В то же время, отмечает Паджет, если бы мы рассматривали исключительно такие характеристики семей флорентийской элиты, как богатство, древность рода и состояния, политический статус, ближайшее окружение, мы бы не об-

[1] *Padgett J. F., Ancell Ch. K.* Robust Action and the Rise of the Medici // American Journal of Sociology. 1993 (May). № 98(6). P. 1259–1319.

наружили существенных различий между семейством Медичи и остальными олигархами. Этот пример показывает перспективность использования методов центральности для углубленного понимания социальных процессов и явлений.

Отметим, что в рамках анализа социальных сетей методы измерения центральности продолжают активно разрабатываться и обсуждаться. Вопросы о том, насколько структурная позиция определяет степень влияния социального субъекта в сети отношений, какой способ измерения этого влияния наиболее эффективен, до сих пор не получили однозначного ответа. В данной области сетевого анализа каждый исследователь может, исходя из своих собственных убеждений, выбрать наиболее подходящий способ измерения центральности.

Централизация сети **(*показатели групповой центральности*).** Централизация сети измеряется посредством ряда групповых показателей центральности, отражающих степень неравенства индивидуальных показателей, что позволяет рассматривать их как аналог дисперсии (меры неоднородности). В частности, к ним относятся *групповые индексы центральности Фримана по степени*, *близости и посредничеству*. Каждый из этих показателей равен сумме отклонений индивидуальных показателей от максимально наблюдаемого, выраженной в процентах от теоретически возможного максимума. Групповые индексы равны нулю, когда все индивидуальные показатели равны между собой, и равны 100, если в графе доминирует одна вершина, т. е. граф сильно централизован (вроде «колеса» с одной центральной осью и периферией). В отличие от дисперсии, групповые индексы не зависят от размера графа.

Вышеназваные показатели сетевого анализа имеют особую значимость при анализе социоцентрических сетей. Для описательного анализа эго-сетей Р. Берт считает более приемлемым использовать такие показатели, как эффективный размер сети и ограниченность[1]. Фундаментальная идея Берта состоит в том, что индивид включен в макроконтекст посредством локальных микросвязей, которые одновременно и ограничивают его, и обеспечивают возможностями доступа к ресурсам. Исследование структуры эго-сети индивида дает возможность понимания мотивов индивидуального поведения, его отличия от поведения других акторов.

Эффективный размер эго-сети представляет собой разность между общим числом связей эго и показателем «избыточности» этих связей. Показатель избыточности напрямую зависит от плотности эго-сети и количества транзитивных триад в ней, т. к. с увеличением плотности сети растет вероятность того, что ее акторы будут предоставлять эго одну и ту же информацию или ресурсы.

---

[1] *Burt R.* Structural Holes Versus Network Closure at Social Capital // Social Capital Theory and Research / ed. by N. Lin, K. Cook, R. Burt. NY, 2001.

Показатель ограниченности эго-сети прямо пропорционален ее избыточности, он определяется через число нетранзитивных (открытых) триад в сети – «*структурных дыр*» в терминологии Берта. Данный показатель основан на предположении, что если эго является единственным источником необходимых ресурсов для своих партнеров (т. е. партнеры эго не связаны между собой), то они не могут существенно влиять на его поведение. Соответственно, чем больше структурных дыр в эго-сети, тем больше конкурентных преимуществ у эго.

Теория «структурных дыр», развиваемая Бертом, проливает свет на то, как социальные сети становятся орудием конкурентной борьбы на внутрикорпоративном и межкорпоративном рынках. Структурные дыры – это разрывы в структуре знакомств внутри сети. На идеальном рынке, где все акторы знают друг друга, информация могла бы распространяться равномерно. В реальном же рынке существуют «дыры»: разделенные ими акторы не знают о тех товарах и услугах, которые они могли бы предложить друг другу. «Дыры» в сети означают, что люди настолько погружены в дела своего круга контактов, что обращают мало внимания на то, что происходит в другом кругу. Здесь и возникает предприниматель, связывающий их друг с другом через себя. Тем самым Берт показывает особую роль знакомств (слабых связей) в условиях несовершенного рынка и ограниченных инвестиций.

Анализ сплоченных подгрупп. С помощью методов анализа сплоченных подгрупп исследователи осуществляют формализацию интуитивных и теоретических определений социальной группы средствами сетевого анализа. Сплоченная подгруппа, или подграф, – это некоторое количество акторов, между которыми существуют сильные, направленные, интенсивные или позитивные связи. В широком смысле сплоченность понимается как единство, общность норм и интересов, взаимные симпатии членов группы. Основными ее индикаторами выступают взаимность и частота контактов акторов, близость и достижимость вершин подграфа. Выделяют два основных подхода к анализу сплоченных подгрупп. Первый заключается в измерении и сравнении структурных характеристик акторов внутри подгруппы, второй – в измерении и сравнении структурных характеристик акторов и внутри подгруппы, и с внешними по отношению к ней акторами. Р. Ханнеман определяет первую группу методов анализа сплоченных подгрупп как подход «*снизу вверх*»[1], имея в виду, что в этом случае объяснение особенностей сетевой структуры осуществляется через определение размеров и числа клик и кликоподобных групп.

Кликой называется связанный подграф, включающий не менее трех узлов, каждый из которых является смежным со всеми остальными; численность (размер) клики обозначается буквой *n*. Другими словами, кли-

[1] *Hanneman R. A., Riddle M.* Указ. соч.

ка – это группа акторов, где каждый связан с каждым и в данную группу не могут быть включены другие акторы, поскольку они не имеют связей со всеми членами клики. На рис. 4.6 изображен граф с двумя кликами: *1–2–5* и *2–3–4–5*.

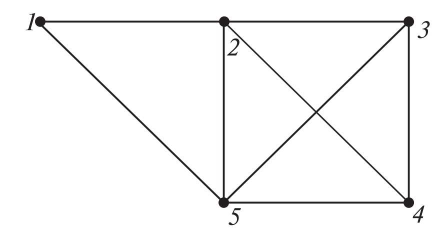


Рис. 4.6. Граф с двумя кликами

Анализ сплоченных подгрупп проводится только на симметричных матрицах связей. Несимметричную матрицу необходимо преобразовать, оставив только симметричные связи.

Пример 4.1 (продолжение)

В графе (на рис. 4.5) всего 7 клик, одна включает 4 вершины, и шесть – по 3 вершины: 1) *Б–Г–Д–Ж*; 2) *Б–В–Д*; 3) *А–В–Д*; 4) *Б–Д–И*; 5) *Б–Д–З*; 6) *А–Д–К*; 7) *В–Д–К*. При этом наибольшая клика, состоящая из четырех акторов, частично пересекается, «накладывается» на остальные, состоящие из трех акторов. Это особенно хорошо видно из табл. 4.8, где фиксируется уровень принадлежности к каждой клике. Вероятность высчитывается как отношение реализованных смежных вершин к максимально возможному количеству смежных вершин в клике.

Таблица 4.8

Уровень принадлежности к клике

Актор	Клика						
	1	2	3	4	5	6	7
А	0,50	0,67	1,00	0,67	0,67	1,00	0,67
Б	1,00	1,00	1,00	1,00	1,00	0,67	0,67
В	0,50	1,00	0,67	0,67	0,67	0,67	1,00
Г	1,00	0,67	0,67	0,67	0,67	0,33	0,33
Д	1,00	1,00	1,00	1,00	1,00	1,00	1,00
Е	0,00	0,33	0,00	0,00	0,00	0,00	0,33
Ж	1,00	0,67	0,67	0,67	0,67	0,33	0,33
З	0,50	0,67	0,67	1,00	0,67	0,33	0,33
И	0,50	0.67	0,67	0,67	1,00	0,33	0,33
К	0,25	0,67	0,67	0,33	0,33	1,00	1,00

Аналогичная информация представлена в табл. 4.9. На диагонали этой матрицы указан размер каждой клики. Можно также проследить, имеет ли определенная клика общие вершины с другой, и если имеет, то сколько. Так, клика *1* имеет по два общих актора с кликами *2*, *3*, *4*, *5* и по одному общему актору с кликами *6* и *7*.

Таблица 4.9

Пересечение клик

Клика	*1*	*2*	*3*	*4*	*5*	*6*	*7*
1	4	2	2	2	2	1	1
2	2	3	2	2	2	1	2
3	2	2	3	2	2	2	1
4	2	2	2	3	2	1	1
5	2	2	2	2	3	1	1
6	1	1	2	1	1	3	2
7	1	2	1	1	1	2	3

Определение клики, приведенное выше, многие аналитики находят избыточно строгим, так как клики, которые являются относительно небольшими по размеру, чаще всего пересекаются друг с другом, что затрудняет их интерпретацию[1]. Несколько более эффективным является выделение кликоподобных групп, к которым относятся *n-клики* и *n-кланы*, основанные на понятиях связности (достижимости) акторов в подгруппе, расстояний между ними; а также *k-плексы* и *k-ядра*, базирующиеся на количестве входящих и исходящих связей акторов[2].

Вместе с тем неустойчивость, уязвимость *n*-клик и *n*-кланов проявляется в их зависимости от расположения вершин графа. Исключение даже одной вершины или ребра из анализа резко отрицательно влияет на весь результат. Методы идентификации *k*-плексов и *k*-ядер дают более реалистичную картину сплоченных подгрупп относительно небольшого размера и устойчивы к случайному отсутствию связей.

Вторую группу методов выделения сплоченных подгрупп в сети Ханнеман обозначает как подход «*сверху вниз*». Данный подход к декомпозиции сети на подгруппы основан на предположении о том, что актор, имеющий определенное количество связей с членами группы, чувствует свою принадлежность к ней несмотря на то, что может не знать многих ее членов или даже большинство из них.

В последнее время группа методов анализа сплоченных подгрупп обогащается новыми интересными и продуктивными разработками, позволяющими анализировать, что происходит со связями группы в момент противостояния разрушающим силам, например, после удаления актора

[1] *Wasserman S., Faust K.* Указ. соч. С. 256.
[2] Там же. С. 263.

(группы акторов) с высоким показателем центральности либо выполняющего функции посредника. Данный подход исключительно важен для анализа коммуникации в организациях и преступных сообществах, так как позволяет выделить из групп равной плотности те, которые более устойчивы к разрушению. Основная задача в этом случае состоит в нахождении актора, диады или триады, удаление которых позволит наиболее серьезно деформировать структуру связей в сети[1].

Позиционный анализ. Методы позиционного анализа позволяют изучать структурное сходство, эквивалентность акторов и паттернов отношений в полных сетях. Основываясь на атрибутивных переменных (пол, должность, образование и др.), социологи довольно часто разбивают совокупность изучаемых объектов на классы (категории, позиции), например, на бедных и богатых. Внутри данных классов объекты социологического исследования максимально похожи друг на друга, что позволяет делать обоснованные предположения относительно их социального поведения. В рамках позиционного анализа социологи также разбивают совокупность изучаемых объектов на классы, но уже на основании не атрибутивных переменных, а структурных. Данный метод анализа очень важен, так как он обеспечивает исследователя аналитическими инструментами для идентификации ролей, выделения их из множества отношений, существующих в сети.

Классы определяются на основании понятия эквивалентности акторов как множества неразличимых акторов, имеющих сходные отношения в сети, *роль* – как тип отношений между акторами и/или позициями. Заметим, что понятие эквивалентности резко отличается от понятия взаимосвязанной подгруппы, так как акторам, занимающим одинаковую позицию, совсем необязательно иметь какие бы то ни было отношения друг с другом. Таким образом, акторы эквивалентны, когда они имеют одинаковые отношения со всеми другими элементами сети, т. е. когда эквивалентны структура и тип взаимодействий данных акторов с другими. Например, клиенты продавца некоторого товара будут иметь очень мало или вообще не будут иметь связей между собой (вследствие этого они не будут сплоченной подгруппой), но все они будут связаны с продавцом, т. е. паттерны взаимодействий продавцов с клиентами будут эквивалентны.

Различают структурную, атоморфную и регулярную эквивалентность. Определение структурной эквивалентности является наиболее строгим: два актора *структурно эквивалентны* только в том случае, если они имеют одинаковую структуру отношений со всеми другими членами сети. С этой точки зрения структурно эквивалентные акторы занимают одинаковые позиции в сети и являются полностью взаимозаменяемыми. Есте-

[1] *Borgatti P. S., Foster P. C.* The Network Paradigm in Organizational Research: A Review and Typology // Journal of Management. 2003. № 29(6). P. 996.

ственно, такие случаи встречаются достаточно редко, и исследователей интересует возможность измерения степени выраженности структурной эквивалентности.

Атоморфная эквивалентность предполагает сходство акторов, если они имеют идентичную структуру отношений с любыми другими членами сети. В данном случае взаимозаменяемыми являются уже не эквивалентные акторы, а эквивалентные подструктуры графа (рис. 4.7).

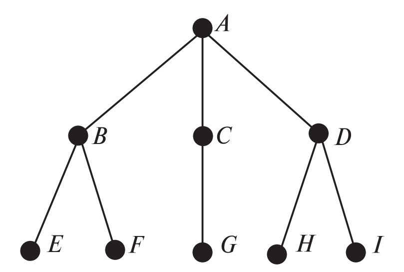


Рис. 4.7. Акторы *B*, *C* и *D* атоморфно эквивалентны

Понятие *регулярной эквивалентности* более универсально: акторы эквивалентны, когда они одинаковым образом взаимодействуют с акторами определенного типа. Так, директор предприятия *X* и директор предприятия *Y* не являются структурно эквивалентными, т. к. они взаимодействуют с сотрудниками разных предприятий. Не являются они и атоморфно эквивалентными, если на предприятии *X* работает в два раза больше сотрудников, чем на предприятии *Y*, но с точки зрения регулярной эквивалентности позиции двух директоров схожи, так как они взаимодействуют с акторами определенного типа – работниками предприятий. Другими словами, социальная позиция и роль определяются в данном контексте взаимозависимо: начальник – подчиненный, мужчина – женщина, муж – жена и др. Данный подход представляется наиболее интересным для социологов, т. к. концепция регулярной эквивалентности наиболее близка к социологической концепции социальной роли.

Итак, мы коротко рассмотрели три уровня позиционного анализа, каждый из которых является менее формальным, чем предыдущий. Структурно эквивалентные акторы могут быть как атоморфно эквивалентны, так и регулярно эквивалентны. Любые атоморфно эквивалентные акторы могут быть регулярно эквивалентными. Но далеко не все регулярно эквивалентные акторы являются структурно или атоморфно эквивалентными.

Стохастический анализ социальных сетей. В последние десятилетия широкое распространение получили методы стохастического анализа социальных сетей, основанные на теории вероятностей и прикладной ста-

тистике. Создание в 2000 г. проекта *www.stocnet.com* и компьютерной программы *STOCNET* позволило в значительной степени ускорить данный процесс. До этого методы статистического анализа были частично доступны в программах *GRADAP*, *UCINET* и др., но они служили лишь дополнением к описательным методам, довольно редко обновлялись и были трудны в практической реализации.

Методы прикладной статистики (меры центральной тенденции, разброса данных, корреляционный, регрессионный анализ, многомерное шкалирование и др.) используются для того, чтобы резюмировать ключевые факты о распределении акторов, атрибутов и отношений. Следует особо отметить, что в рамках сетевых исследований основная задача состоит в анализе отношений не между переменными, а между акторами, т. е. исследователь работает не с распределением характеристик акторов, а с распределением *отношений* между ними.

Методы вероятностного анализа, учитывающие информацию о распределении структурных характеристик в генеральной совокупности сетей с заданными свойствами, предоставляют возможность проверить предположения о форме распределений, значимости параметров модели, пригодности модели для описания данных. Большинство методов теории статистического вывода, традиционно используемых социологами (стандартная ошибка, проверка гипотез), не могут быть применены непосредственно к анализу сетевых данных, т. к. единицы анализа – акторы и отношения между ними – являются взаимозависимыми. Это обстоятельство долгое время сдерживало развитие и применение стохастических методов в обработке сетевых данных.

Прорыв в данной области исследований связан с разработкой статистических моделей экспоненциальных случайных графов – так называемых *p*-моделей*, основанных на марковских случайных графах. Основная идея данных вероятностных моделей состоит в том, что каждая социальная сеть может быть рассмотрена как реализация $x = \{x_{i,j}\}$ случайного двумерного бинарного массива X. Так как элементы массива X являются зависимыми случайными величинами, то можно анализировать структуру зависимостей между соответствующими акторами социальной сети, находить вероятности существования определенных реализаций социальной сети и получать оценки ее параметров. P^*-модели позволяют оценивать сложные статистические модели, в которых связь между акторами рассматривается как зависимая переменная, а структурные свойства социальной сети (баланс, кластеризация, транзитивность и др.) – как независимые.

В итоге можно сделать вывод о том, что сетевой анализ представляет собой не просто набор исследовательских методик и стратегий, а единое методологическое направление, основанное на четких концептуальных принципах, позволяющее решать задачи описания, моделирования и объяснения самых различных аспектов функционирования социаль-

ной структуры общества. Его основное отличие от «классической» исследовательской методологии заключается в том, что единицей наблюдения и анализа являются не атрибуты, ценности или убеждения социальных субъектов, а связи между ними. Формальные категории сетевого анализа, такие как сеть, социальная роль, позиция, группа, клика, популярность, изолированность, престиж, влияние и многие другие, имеют четкую эмпирическую концептуализацию, в связи с чем особенно эффективно используются для репрезентации социальной структуры. Экспликация математических показателей структурных характеристик многократно усиливает социальное исследование, обеспечивая его четкими определениями социальных концептов и облегчая разработку надежных моделей социальных процессов и объяснительных теорий.

Сетевые аналитики имеют уникальную возможность проводить свои исследования одновременно и на индивидуальном, и на групповом уровне, перекидывая тем самым «мост через пропасть» между микро- и макросоциологией. С одной стороны, сущность группы детерминирует взаимоотношения акторов в ней. С другой стороны, тип и характеристики акторов определяют групповую структуру взаимодействий. Поэтому индивидуальное и групповое объяснения не могут быть полностью поняты в отрыве друг от друга. Таким образом, способ, которым сетевые аналитики концептуализируют социальную структуру, является одновременно и общим, и конкретным.

Важно понимать, что сетевой подход не ограничивается описанием связей акторов в обществе. Его основная цель – объяснение, хотя бы частичное, поведения сетевых элементов и системы в целом, апеллирующее к специфическим особенностям взаимодействий между элементами. Как отмечает Д. Ноук, «структура отношений между акторами и расположение индивидов в сети имеет важные поведенческие и аттитюдные последствия как для индивидуальных объединений, так и для всей системы в целом»[1]. Таким образом, сетевой подход исследует ограничения и возможности паттернов взаимодействий по отношению к составляющим его акторам, благодаря чему появляется возможность установить относительно прочную связь между теорией и эмпирией, логикой и опытом.

Самостоятельная работа

Проведите реконструкцию и анализ эго-сети.

1-й шаг. Заполнение таблицы. Вспомните 35 человек, которых Вы знаете лично, тех, кого Вы можете узнать в лицо или по имени и с кем Вы можете вступить в контакт, если возникнет такая необходимость. Можно использовать условные имена (псевдонимы). Впишите эти имена в соответствующий столбец. В следующих столбцах укажите пол и приблизительный возраст каждого человека. На-

[1] *Knoke D.* Political Networks: The Structural Perspective. NY, 1990. P. 13.

пишите своими словами, откуда Вы знаете его/ее (т. е. что Вас связывает – родственные отношения, совместная работа или учеба, дружба, соседство, дружба с одними и теми же людьми, случайное знакомство, связь через посредство другого человека и т. д.).

№	«Имя»	Пол	Возраст	Откуда Вы его/ее знаете, кем он/она Вам приходится
1				
2				
...				
40				

2-й шаг. Заполнение социоматрицы. Внесите имена указанных Вами людей в названия столбцов и строк. Для удобства работы распечатайте сначала таблицу, в которой будет 35 строк и столбцов, а уже потом «с бумаги» внесите в программу UCINET. Для каждой пары людей оцените, знают ли они друг друга, используя следующие значения: «0» – не знают друг друга, «1» – знают друг друга; в результате полученная матрица данных должна быть симметричной.

Пример социоматрицы:

№	«Имя»	Саша	Антон	Петр Петрович	Рыжий	...	Семеныч
1	Саша						
2	Антон						
3	Петр Петрович						
4	Рыжий						
5	...						
6	Семеныч						

3-й шаг. Визуализация и анализ полученной сети. Создайте граф связей с помощью одной из программ визуализации UCINET. Вычислите в этой же программе: 1) показатели плотности, связности сети; 2) показатели центральности по степени, близости и посредничеству для акторов полученной сети; 3) выполните анализ сплоченных подгрупп с помощью выделения в сети *n*-клик (при n = 5 и n = 7).

Литература

Батыгин, Г. С. Сетевые взаимосвязи в профессиональном сообществе социологов: методика контент-аналитического исследования / Г. С. Батыгин, Г. В. Градосельская // Социол. журн. 2001. № 1.

Габышева, Л. К. О некоторых концепциях сетевого моделирования / Л. К. Габышева // Социология : 4М. 2008. № 27.

Градосельская, Г. В. Сетевые измерения в социологии : учеб. пособие / Г. В. Градосельская. М., 2004.

Грановеттер, М. Экономическое действие и социальная структура: проблема укорененности / М. Грановеттер // Эконом. социология [Электронный ресурс]. М., 2002. Т. 3, № 3.

Лаврусевич, П. Е. Социальные сети в стратегиях трудоустройства на российском рынке труда / П. Е. Лаврусевич // Эконом. социология [Электронный ресурс]. М., 2006. Т. 7, № 2.URL: http://ecsoc.hse.ru/issues/2006-7-2/index.html.

Морено, Я. Л. Социометрия: Экспериментальный метод и наука об обществе / Я. Л. Морено. М., 2004.

Пауэлл, У. Сети и хозяйственная жизнь / У. Пауэлл, Л. Смит-Дор // Эконом. социология [Электронный ресурс]. 2003. Т. 4, № 3.URL: http://www.hse.ru/mag/ecsoc/2003-4-3/26591870.html.

Печенкин, В. В. Методы анализа социальных сетей на примере визуализации структуры предпочтения профессий / В. В. Печенкин // Социология : 4М. 2001. № 13.

Сивуха, С. В. Социальная сеть общественных организаций как форма социального капитала / С. В. Сивуха // Социология. 2003. № 4.

Molina, J. L. The Informal Organizational Chart in Organizations: An Approach from the Social Network Analysis / J. L. Molina // Connections [Electronic jornal]. 2001. № 24 (1). URL: http://www.insna.org/PDF/Connections/v24/2001_I-1_78-91.pdf.

Глава 5

КОГОРТНЫЙ АНАЛИЗ

Когортный анализ – это сравнительный анализ изменений в социальном поведении одной или нескольких когорт с течением времени. В демографическом контексте, где возник термин «когорта», он используется для обозначения совокупности людей, у которых в один и тот же период времени произошло определенное *демографическое событие*. Например, когортой является группа лиц, родившихся в 1961–1965 гг., или группа женщин, вступивших в первый брак в 1980 г. Интервал времени, выбираемый для выделения когорты (один год, 5 лет и т. п.), зависит от целей анализа и особенностей исходных данных. Исторически первым видом когортного анализа был «*продольный*» демографический анализ, направленный на изучение частоты событий, происходящих в когорте в зависимости от ее «возраста», т. е. промежутка времени между ее образованием и изучаемым событием. В качестве событий могут рассматриваться вступление в брак, развод, рождение первого ребенка, смерть и т. п.

В настоящее время когортный анализ как аналитический метод используется не только в демографии, но и в других отраслях знания. Например, в социологии широкое распространение получили лонгитюдные исследования жизненного пути образовательных когорт, выделенных по критерию завершения определенного уровня образования в определенном году[1]. В качестве когорты может рассматриваться и группа предприятий, созданных в один и тот же год или в ограниченный временной период (например, в годы перестройки), автомобили одной модификации, вина урожая одного года и т. п.

Методологические и функциональные возможности метода постоянно расширяются. Если «продольный» анализ изучал одну когорту и частоту происходящих в ней событий, то современный когортный анализ позволяет сравнивать жизненные пути нескольких когорт, а также возрастные и временные аспекты изучаемых процессов. В любом случае исследова-

---

[1] См., напр.: Социальное расслоение возрастной когорты. Выпускники 80-х в постсоветском пространстве / отв. ред. М. Титма. М., 1997.

ние с использованием этого метода является достаточно продолжительным (сопоставимым по длительности с «жизнью» изучаемых объектов), в значительной мере ретроспективным (может быть осуществлено не ранее, чем изучаемое событие наступит хотя бы для части изучаемой когорты) и методологически сложным, т. к. предполагает разграничение эффектов возраста, когорты и времени.

Основная методологическая проблема состоит в том, что возраст – чрезвычайно коварная объясняющая (независимая) переменная. Когда мы осуществляем сравнительный анализ по возрасту, возникает явление, получившее в экспериментальных исследованиях название смешения эффектов нескольких факторов. В случае когортного анализа это эффекты трех факторов – возраста, когорты и времени. *Фактор возраста* соотносится с такими свойствами личности, как физическое здоровье, степень мобильности, интерес ко всему новому и способность его восприятия и т. д. *Фактор когорты* связан, в первую очередь, с условиями ее возникновения, становления, социализации, жизненного опыта, доступа к различного рода ресурсам и т. п. Например, пожилые люди значительно хуже относятся к телевизионной рекламе, чем молодые, не только потому, что нелюбопытны и им не нравится агрессивный стиль рекламных роликов (эффект возраста), но также потому, что вся прожитая жизнь приучила их к экономии и осторожности в принятии решений о покупках (эффект когорты). *Фактор времени* отражает перемены, происходящие в макросреде, – экономические, политические, социальные, информационные и др. Например, с помощью последовательной информационной политики, разъясняющей необходимость рекламы для развития экономики, можно добиться снижения неприятия рекламы во всех возрастных когортах.

Пример 5.1. Надежность автомобилей

Замечательный пример того, как опасно строить на данных единовременных исследований («срезов») серьезные выводы и прогнозы, приводит Р. Дэвис в статье «От срезового к лонгитюдному анализу»[1]. Представим себе, что сервисный центр ведет статистику надежности стоящих на обслуживании автомобилей (индикатор – число дней простоя в год) с группировкой по возрастам (табл. 5.1).

Таблица 5.1

Число дней простоя в зависимости от возраста автомобиля

Возраст автомобиля (число лет)	1	2	3	4	5
Среднее число дней простоя в год	4	3	15	16	18

[1] Источник: *Davis R. B.* From Cross-Sectional to Longitudinal Analisys // Analyzing Social & Political Change: A Casebook of Methods / ed. by A. Daleand, R. B. Davis. London, 1994. P. 24.

Наиболее простым объяснением этой вполне очевидной зависимости является эффект возраста: чем старше автомобиль, тем менее он надежен. Если это так, то можно предположить, что в последующие годы автомобили будут «вести себя» аналогично, бизнес можно планировать как устойчивый, и рассчитывать на тот же уровень прибыли. Но специалисту, знакомому с эффектами когорты, столь же очевидным представляется и *альтернативное объяснение*: два года назад сменилось поколение автомобилей. Новые автомобили (которым в данный момент 1–2 года) будут и впредь вести себя надежно, автомобили старшего поколения (3 года и старше) изначально не были достаточно надежными и таковыми и останутся. Поскольку с каждым годом их будет становиться меньше, для поддержания бизнеса (не говоря уже о его развитии) понадобятся «свежие» решения.

Понять, с каким из случаев мы имеем дело, по данным одного «среза» невозможно. Необходим сравнительный анализ трендов для автомобилей каждого года выпуска, т. е. когортный анализ, как минимум, за последние пять лет.

Чтобы осуществить его, имеющуюся статистику следует организовать в виде таблицы, строки которой будут соответствовать возрасту автомобилей, а столбцы – годам наблюдения. В клетке таблицы, образованной пересечением строки и столбца, указывается усредненное значение изучаемого показателя для соответствующей когорты в соответствующем году. Такую таблицу читают по диагонали, по мере «старения» когорты и течения времени.

Пример 5.1 (продолжение)

В нашем примере данные, соответствующие гипотезе об эффекте возраста, могли бы выглядеть, как показано в табл. 5.2. Когорта автомобилей выпуска 1986 г. в первый год работы простаивала в среднем 4 дня, во второй (1987) – 3 дня, на третьем году эксплуатации (1988) надежность резко снизилась – 17 дней простоя – и с тех пор остается на том же уровне: в 1989 г. – 15 дней, в 1990 г. – 18. Аналогично можно проанализировать «поведение» автомобилей 1987 и последующих годов выпуска. Заметим, что столбец, соответствующий 1990 г., полностью совпадает с распределением из табл. 5.1.

Таблица 5.2

Число дней простоя в зависимости от возраста автомобиля (эффект возраста)

Возраст автомобиля	Календарный год					Когорта (год образования)
	1986	1987	1988	1989	1990	
5					**18**	**1986**
4				**15**	16	1987
3			**17**	14	15	1988
2		**3**	4	4	3	1989
1	**4**	3	3	3	4	1990

Если же верна гипотеза об эффекте когорты, то взятые в ретроспективе данные могут выглядеть, например, так, как показано в табл. 5.3. Заметим, что последний столбец также совпал с данными из табл. 5.1.

Таблица 5.3

Число дней простоя в зависимости от возраста автомобилей (эффект когорты)

Возраст автомобиля	Год выпуска					Когорта (год образования)
	1986	1987	1988	1989	1990	
5					18	1986
4				15	16	1987
3			16	17	15	1988
2		14	18	15	**3**	**1989**
1	17	15	16	**4**	**4**	**1990**

Из табл. 5.3 видно, что когорты автомобилей 1986, 1987 и 1988 гг. выпуска с первого года эксплуатации показывали ту же надежность, что и в текущем 1990 г., а именно 14—17 дней простоя в год. Автомобили когорт 1989 и 1990 гг. выпуска изначально имеют другую надежность, и нет никаких оснований предполагать, что в возрасте 3—5 лет они будут столь же часто нуждаться в ремонте, как автомобили предшествующего поколения. Как, впрочем, пока нет данных, свидетельствующих о том, что их надежность с возрастом не уменьшится. Поэтому исследование необходимо продолжать.

Таким образом, когортный анализ базируется на специальной форме представления статистических данных за несколько лет. Он «пришел» из демографии, где возрастные когорты традиционно сравниваются, например, по продолжительности жизни и другим показателям, полученным с помощью текущего статистического учета и в ходе переписей населения. Рассмотренный выше пример из статьи Дэвиса также базируется на статистике, точнее, на внутренней статистике фирмы.

Однако применение метода когортного анализав в социологии и маркетинге было бы чрезвычайно ограничено, если бы он мог осуществляться только на полных данных о генеральной совокупности. Метод в простейшем его виде может быть реализован и на данных регулярных дескриптивных обследований по репрезентативной выборке, которая в ходе обработки данных разделяется на возрастные группы, и для каждой группы вычисляется среднее арифметическое изучаемого показателя. В этом случае условиями корректности сравнений являются относительно небольшое значение дисперсии, неизменность генеральной совокупности и методов измерения «трендовых» показателей, а также схемы формирования выборки. Кроме того, при обнаружении различий необходимо проверять их статистическую значимость.

Пример 5.2. Потребление безалкогольных напитков

Ретроспективный когортный анализ данных выборочного повторного исследования был предпринят в начале 1980-х гг. Дж. Ренцом, Ф. Рейнолдсом и Р. Стаутом с целью опровергнуть бытовавший в то время среди маркетологов стереотип, согласно которому употребление безалкогольных напитков с возрастом уменьшается[1]. Анализу были подвергнуты данные четырех маркетинговых опросов (1950, 1960, 1970 и 1980), в каждом из которых респондентов спрашивали, употребляют ли они безалкогольные напитки. Выборка по каждому из опросов была сгруппирована по возрасту в 10-летние интервалы. В качестве трендового показателя для возрастной группы определялся процент респондентов, ответивших положительно (табл. 5.4).

Таблица 5.4

Распределение опрошенных по потреблению безалкогольных напитков(% от объема возрастной группы)

Возраст	Год опроса			
	1950	1960	1970	1980
18–19	52,9	62,6	73,2	81,0
20–29	45,2	60,7	76,0	75,8
30–39	33,9	46,6	67,7	71,4
40–49	23,2	40,8	58,6	67,8
50–60	18,1	28,8	50,0	51,9

По форме представления табл. 5.4 напоминает план эксперимента с двумя независимыми переменными (факторами) – возраст и год исследования. На первый взгляд, гипотеза об эффекте возраста (снижении с возрастом потребления безалкогольных напитков) подтверждается в каждом из четырех опросов. Очевиден также эффект времени, проявляющийся в том, что от опроса к опросу процент потребителей безалкогольных напитков растет в каждой возрастной группе.

С целью «продольного» когортного анализа читать ту же таблицу следует по диагонали, по мере течения времени и старения когорты. Каждая полная или неполная диагональ соответствует одной из когорт, выделяемых в процессе анализа. Табл. 5.5 получена из табл. 5.4 после выделения и обозначения когорт. Например, когорта № 4, данные о которой выделены в табл. 5.5 жирным шрифтом, в 1950 г. находилась в возрасте 20–29 лет и, соответственно, родилась в 1921–1930 гг. В 1950 г. 45,2 % опрошенных представителей этой когорты ответили, что они употребляют безалкогольные напитки. В 1960 г., достигнув возраста 30–39 лет, в этой же когорте безалкогольные напитки употребляли 46,6 % опрошенных, в 1970 г., в возрасте 40–49 лет – 58,6 %; в 1980 г. в возрасте 50–60 лет – 51,9 %.

[1] Источник: *Малхотра Н. К.* Указ. соч. С. 144.

Таблица 5.5

Потребление безалкогольных напитков: «продольный» когортный анализ (%)

Возраст	Год опроса				Номер когорты
	1950	1960	1970	1980	
18–19	52,9	62,6	73,2	81,0	
20–29	**45,2**	60,7	76,0	75,8	8 (1961–1970)
30–39	33,9	**46,6**	67,7	71,4	7 (1951–1960)
40–49	23,2	40,8	**58,6**	67,8	6 (1941–1950)
50–60	18,1	28,8	50,0	**51,9**	5 (1931–1940)
Номер когорты		1 (1891–1900)	2 (1901–1910)	3 (1911–1920)	**4 (1921–1930)**

Разделение выборки каждого из четырех опросов на 10-летние возрастные группы позволяет выделить в этой таблице 8 возрастных когорт, наблюдение за которыми имело разную продолжительность. Так, когорта № 1 (1891–1900 гг. рождения) наблюдалась только в исследовании 1950 г. в возрасте 50–60 лет. Ее, конечно, нельзя подвергнуть «продольному» анализу, однако она обладает безусловной ценностью для сравнений с последующими когортами по мере достижения ими возраста 50–60 лет, а также для сравнения с другими возрастами в «срезовом» исследовании 1950 г. Аналогичным образом дело обстоит и с когортой № 8 (1961–1970 гг. рождения), представители которой впервые появляются в опросе 1980 г. Остальные 6 представленных в таблице когорт позволяют проследить изменение потребительского поведения в отношении безалкогольных напитков на протяжении 10–30 лет (в зависимости от числа последовательных опросов, в которых они принимали участие). Четыре когорты (под номерами 2, 3, 5 и 7) убедительно показывают, что сформированная в том или ином возрасте привычка пить безалкогольные напитки с течением времени не исчезает и число потребителей таких напитков в когорте растет. Исключение составляют когорта № 4, в которой снижается потребление безалкогольных напитков при переходе из возрастной группы сорокалетних (58,6 %) в группу пятидесятилетних (51,9 %), а также когорта № 6, в которой аналогичная тенденция наблюдается при переходе из группы двадцатилетних (76 %) в группу тридцатилетних (71,4 %). Узнать, являются эти отклонения следствием ошибки выборки или предвещают появление некой новой тенденции, можно только продолжив исследования с прежней или даже более короткой периодичностью.

Таким образом, читая таблицу когортного анализа *по диагонали*, мы получаем данные о поведении когорты на протяжении определенного временного периода. Анализ таблицы *по столбцам* позволяет провести сравнение потребительского поведения различных возрастных групп в конкретный момент времени, например, с целью сегментирования рынка потребителей. *Строки таблицы* дают возможность проследить, как менялось с годами поведение каждой возрастной группы. Таким образом, когортный анализ позволяет разделить эффект когорты (по диагонали),

эффект возраста (по столбцам) и эффект времени или социокультурных изменений (по строкам).

Количество замеров, необходимых для когортного анализа, зависит от объекта исследования, а также от того, насколько быстро протекают изучаемые процессы. Нередко даже относительно непродолжительные, но систематические наблюдения позволяют эффективно использовать метод.

Пример 5.3. Распределение интернет-аудитории по возрасту

В табл. 5.6 представлены результаты когортного анализа белорусской интернет-аудитории по данным Независимого института социально-политических и экономических исследований[1].

Таблица 5.6

Распределение интернет-аудитории по возрасту (доля пользователей в % в возрастной группе)

Возраст	Год		
	1998	2003	2008
16–19	3,3	44,1	71,0
20–24	4,5	43,4	75,7
25–29	**8,0**	33,8	61,3
30–34	2,4	**23,3**	48,5
35–39	2,4	17,7	**48,6**
40–44	1,1	17,7	43,2
45–49	1,1	11,9	36,8
50–54	0,5	9,8	30,3
55–59	0,5	2,9	10,5
60+	0,3	2,8	3,6
Доля по выборке	2,3	17,8	31,2
Объем выборки	1474	1488	1531

Когортный анализ, в частности, показал, что первые интернет-пользователи (в заметном количестве) появились в Беларуси в 1998 г. среди наиболее социально и экономически активной группы населения 25–29-летнего возраста. В 2008 г. в этой когорте, достигшей 35–39-летнего возраста, пользователем Интернета является каждый второй. Она сохраняет преимущество перед старшими поколениями, что можно идентифицировать как эффект возраста. В то же время эта группа уступает более молодым, что, безусловно, является эффектом когорты, т. к. они изначально находились в более выгодных условиях, во-первых, в силу многократно-

[1] *Терещенко О. В.* Когортный подход к анализу белорусской аудитории интернета // Тезисы III Всероссийского социологического конгресса (Москва, 21–24 октября 2008 г.). М., 2008.

го повышения экономической и технической доступности Интернета, во-вторых, благодаря образовательной политике государства.

Основным недостатком представленных результатов является отсутствие статистической значимости различий между соседними возрастными группами внутри каждого временного периода. Это вызвано, прежде всего, использованием коротких 5-летних возрастных интервалов, что обусловлено не только краткостью периода наблюдения, но и стремительностью развития процесса «интернетизации». С другой стороны, нами использованы данные, которые собирались не с целью последующего когортного анализа, а для опросов общественного мнения. Выборка объемом 1500 респондентов вполне репрезентативно представляет десятимиллионное население Беларуси в целом, но ее не достаточно для сравнительного анализа десяти 5-летних возрастных групп.

Эффект когорты нейтрализуется применением лонгитюдных исследований. Лонгитюд – это панельное исследование, в котором выборка (*панель*) сформирована по критерию возраста или завершения определенной ступени образования. Этот подход предназначен для изучения жизненных путей возрастной когорты. Он ценится за уникальную возможность изучения изменений в поведении респондентов, их ценностей и установок под влиянием тех или иных внешних событий. Основным недостатком лонгитюдных исследований является высокая стоимость и трудоемкость, связанная с необходимостью поддерживать контакты с респондентами на протяжении многих лет и обеспечивать их участие в очередных этапах. Кроме того, систематические ошибки выборки, допущенные на первом этапе, даже если впоследствии они были исправлены, оказывают влияние на все последующие этапы лонгитюдного исследования. На репрезентативность поздних этапов лонгитюда влияет эффект «вымирания панели», т. е. потеря значительного числа респондентов по причинам миграции, смертности, отказов от дальнейшего участия и т. д.

Сравнительный «продольный» анализ нескольких когорт может осуществляться как на основе данных параллельных лонгитюдов, идущих со «сдвигом» в несколько лет, так и данных мониторинга – серии повторных исследований с независимыми выборками. Мониторинг значительно дешевле и организационно проще, чем лонгитюд, и построенные на его основе тренды менее подвержены систематическим ошибкам, однако он не обладает столь мощным аналитическим потенциалом.

Пример 5.4. Изменение потребительских предпочтений

Преимущество панельных исследований по сравнению с мониторинговыми Н. К. Малхотра демонстрирует следующим примером из маркетинговых исследований[1]. В табл. 5.7 представлены результаты двух последовательных опросов, на основе которых можно заключить, что в промежуток между опросами ничего не изменилось, все торговые марки сохранили своих потребителей, следовательно, потребители удовлетворены и лояльны.

[1] Источник: *Малхотра Н. К.* Указ. соч. С. 147.

Таблица 5.7

Распределение результатов двух последовательных опросов

Торговые марки	Первый опрос	Второй опрос
Товар марки *А*	200	200
Товар марки *Б*	300	300
Товар марки *В*	500	500
Всего	1000	1000

Аналитические возможности мониторинга на этом исчерпаны. Но если исследование было панельным, его данные могут быть дополнительно подвергнуты перекрестной табуляции (табл. 5.8). В этом случае становится очевидным, что сохранившаяся численность потребителей не свидетельствует о сохранении их состава и, соответственно, удовлетворенности потребляемой маркой. По данным табл. 5.8 можно проанализировать не только распределение потребительских предпочтений, но и отношение к товарам конкурентов. При этом становится очевидным перемещение потребителей между брендами – и тех, кто недоволен маркой (отток), и тех, кто недоволен конкурирующей маркой (приток).

Таблица 5.8

Перекрестная классификация данных двух панельных опросов

Марки товаров (первый опрос)	Марки товаров (второй опрос)			Всего
	А	*Б*	*В*	
А	100	50	50	200
Б	25	100	175	300
В	75	150	275	500
Всего	200	300	500	1000

Кроме того, специфика панели (наличие контактной информации с респондентами) предоставляет дополнительную возможность выбрать респондентов, сменивших марку, и провести с ними глубинные интервью или фокус-группы для уточнения причин изменения потребительских предпочтений.

В организациях, регулярно проводящих массовые опросы, есть все условия для систематического использования когортного анализа. Просто необходимо выделять в текущих исследованиях вопросы, перспективные для образования трендов; тщательно тестировать их формулировки; повторять их время от времени, при удобном случае, по возможности на выборках большого объема; стараться не изменять формулировки, а при крайней необходимости изменений учитывать их влияние на долгосрочные тренды. Наконец, тривиальный совет: не группировать возраст в анкете, ибо группировка, заложенная в текст анкеты, ограничивает возможность последующего выделения когорт.

Самостоятельная работа

Проанализируйте динамику белорусской аудитории FM-радиостанций по времени, по возрасту и по когортам (табл. 5.9)

Таблица 5.9

Аудитория FM-радиостанций (% по возрастной группе)[1]

Возраст	Год			
	1996	2001	2006	2011
18–19	18,7	75,8	69,4	57,7
20–24	17,3	72,3	71,0	62,5
25–29	18,7	66,4	66,7	54,5
30–34	12,7	64,6	67,2	55,0
35–39	8,8	46,7	59,0	57,2
40–44	4,7	45,2	59,7	49,2
45–49	5,2	42,6	51,8	54,5
50–54	3,3	27,3	40,6	50,8
55–59	2,6	21,7	41,2	43,0
60–64	0,0	16,7	25,6	40,4
65–69	0,0	8,8	21,9	28,0
70+	0,0	8,6	16,8	11,2
Всего	11,3	43,0	49,6	47,0

Литература

Барашкова, А. С. Когортный анализ поведения населения на брачном рынке / А. С. Барашкова // Региональная экономика: теория и практика. 2011. № 15 (198).

Науэн, М. С. Метод когортного анализа в социологии / М. С. Науэн // Журн. социологии и социальной антропологии. 2006. Т. 9, № 3.

Терещенко, О. В. Метод когортного анализа в социальных исследованиях / О. В. Терещенко // Социология : 4М. 2009. № 29.

[1] Источник: НИСЭПИ [Электронный ресурс]. 2011. URL: http://www.iiseps.org.

ЗАКЛЮЧЕНИЕ

Анализ данных – важный, достаточно длительный и при этом увлекательный этап социальных исследований. Его успех определяется целым рядом факторов, среди которых не последнее место занимают профессионализм и опыт аналитика, а также знание им предмета исследования. В то же время существуют общие принципы выбора и применения методов анализа данных, соблюдение которых позволяет значительно повысить их эффективность. Важнейшими из этих принципов являются требование идти «от задачи» и комплексное применение методов анализа данных.

Принцип «*от задачи*» означает не только абсолютный приоритет задач исследования при выборе методов анализа данных, но также обеспечение возможности их применения на этапе разработки программы исследования. Связано это, в первую очередь, с ограничениями, которые применение метода накладывает на исходные данные. Уже при постановке задач исследования необходимо выбрать методы, которыми они будут решаться, и разрабатывать инструментарий, а в некоторых случаях и стратегию формирования выборки, с учетом соответствующих ограничений.

Еще один принцип заключается в том, чтобы при решении исследовательских задач не ограничиваться применением единственного метода. *Комплексное применение* методов анализа данных диктуется не только сложностью аналитического процесса, но и возможностью получить более обоснованные и надежные результаты. Существуют две основные стратегии комплексного применения статистических методов в социологических исследованиях: 1) последовательное применение нескольких методов на разных этапах решения задачи; 2) параллельное применение нескольких методов на одном этапе решения задачи.

Последовательное использование методов может быть обусловлено: 1) сложностью задачи и многоэтапностью ее решения; 2) необходимостью предварительной проверки некоторых условий или преобразования данных с помощью других методов; 3) возможностью применения некоторых методов при содержательной интерпретации полученных результатов.

Исследовательские задачи, как правило, решаются в несколько этапов: анализ одномерных распределений; проверка гипотез (как дедуктивных, выдвинутых при планировании исследования, так и индуктивных, возникающих по мере «погружения» исследователя в решение задачи); анализ парных связей между переменными; построение моделей, позволяющих решать задачи снижения размерности, классификации, изучать множественные причинные связи.

Параллельное использование статистических методов при решении социологической задачи позволяет: 1) обосновать выбор метода, модель которого наиболее адекватно описывает изучаемое явление; 2) реализовать более глубокий анализ данных с учетом возможностей и ограничений каждого метода; 3) осуществить проверку надежности полученных результатов посредством их сравнительного анализа.

При проверке надежности сходство результатов, полученных разными методами, может служить подтверждением того, что найденная закономерность действительно существует. Различие рассматриваемых результатов, согласно меткому наблюдению Э. А. и Р. Э. Абгарянов[1], может свидетельствовать о наличии одной из следующих ситуаций: 1) искомая закономерность в действительности не существует, и полученные результаты являются артефактными; 2) искомая закономерность существует, но только одна из построенных моделей ей соответствует; 3) искомая закономерность существует, но ни одна из построенных моделей ей не соответствует. Для определения, какая именно ситуация имеет место, исследователю требуются профессиональный опыт, владение теорией изучаемого явления, а также глубокое понимание используемых методов и моделей.

Работы российских и белорусcских социологов, помещенные в приложении, являются образцами аналитического мастерства. В них можно найти примеры как последовательного, так и параллельного применения методов анализа данных при решении содержательных задач.

[1] *Абгарян Э. А., Абгарян Р. Э.* Проблемы математизации социологических исследований. М., 1983. С. 86.

ПРИЛОЖЕНИЕ

СТРУКТУРА ЛИЧНОСТНОГО ОБРАЗА НАРОДНОГО ДЕПУТАТА В СОЗНАНИИ ИЗБИРАТЕЛЕЙ Г. МОСКВЫ[1]

Ю. Л. Качанов, И. В. Задорин

Постановка задачи

До последнего времени советские социологи практически не имели возможности изучать объекты, так или иначе связанные с политической структурой общества и, вообще, с реальной политической жизнью. Поэтому понятен тот повышенный интерес, который вызвали в их среде первые за многие годы свободные и демократические выборы.

Многочисленные публикации, появившиеся по итогам избирательной кампании 1989 г., кроме результатов исследований, продемонстрировали весьма различные подходы к изучению советского электората[2] [1–3]. Для нас наибольший интерес представляют те работы, где предпринимаются попытки не только проанализировать мнения избирателей о кандидатах в народные депутаты и их предвыборных программах, но и выявить основания выбора и мотивы тех или иных предпочтений, скрытые порой и от самого избирателя.

Как показал опыт первых опросов, можно выделить четыре основных фактора, определяющих выбор избирателей, а именно: социально-профессиональный статус кандидата, его политическая декларация (программа), партийность и, наконец, его личностные качества (точнее, тот личностный образ, который кандидат создает в сознании избирателя). Судя

---

[1] Опубликовано: Демократические институты в СССР: проблемы и методы исследования / сост. И. В. Задорнов ; науч. ред. О. М. Маслова. М. : ЦИРКОН, 1991. С. 28–43.

[2] Электорат – полная совокупность избирателей какого-либо округа, региона или страны.

по некоторым исследованиям [2], все более решающим в процессе выбора становится именно личностный фактор. Вместе с тем изучение механизма восприятия избирателями кандидата в депутаты с этой стороны является и наиболее трудным делом в силу большой размерности личностного образа (значительного количества возможных качеств, составляющих образ), существенно большей, нежели у других вышеназванных факторов.

Известно, однако, что человек в ситуации выбора (любого) старается редуцировать свою задачу, уменьшив число факторов, определяющих решение (т. е. размерность пространства оснований выбора), ограничив его, как правило, тремя-пятью. Более того, люди, чаще всего, априори уже имеют порожденную их деятельностью и соответствующим жизненным опытом систему категорий, критериев, эталонов, через которую они осуществляют восприятие, оценивание и выбор, иными словами, они имеют социальный стереотип, соответствующий ситуации выбора.

Согласно У. Липпману, стереотипы — это упорядоченные, схематичные, детерминированные культурой «картинки» мира в сознании человека, экономящие его усилия при восприятии социальных объектов и защищающие его ценности и позиции. В советской литературе [4] социальный стереотип определяется как эмоционально окрашенный, схематический, стандартный образ, для которого характерны поляризация оценки, жесткая фиксация, интенсивная аффективная коннотация (оценивание), устойчивость. Стереотипы несут функцию социальных эталонов, образцов, фиксирующих опыт.

Задача нашего исследования — изучение социальных стереотипов восприятия личностного образа народных депутатов.

В основе применявшихся экспериментов лежит измерение среднестатистических оценочных реакций, свойственных выборочной совокупности респондентов. Данные измерений обрабатывались методами альфа-факторного анализа [5] и многомерного шкалирования [6]. В последнем случае использовался статистический пакет CLAMS (разработчик — СП «Динамика»).

Исследование проводилось в ноябре — декабре 1989 г. в преддверии избирательной кампании 1990 г.[1] Генеральной совокупностью выступало множество избирателей г. Москвы. Модель выборки строилась как простая случайная; генеральная совокупность репрезентирована по значению для респондентов фиксированного набора личных качеств депутата. Ошибка репрезентативности данной выборки с доверительной вероятностью 0,954 не превышает 0,0555. (Метод расчета выборки см. в [7]). Объем выборочной совокупности составил 1431 человек, ее социально-демографическая структура представлена на рис. 1—4.

[1] Авторы работы хотели бы выразить благодарность руководству Московской ассоциации центров НТТМ, при финансовой поддержке которой проводилось настоящее исследование, а также Д. В. Захарову, В. К. Коробову, М. Г. Пугачевой и Н. А. Шматко, непосредственно принимавшим в нем участие.

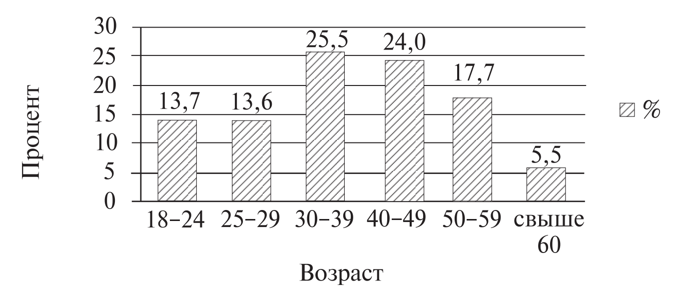


Рис. 1. Распределение выборочной совокупности по возрасту



Рис. 2. Распределение выборочной совокупности по полу

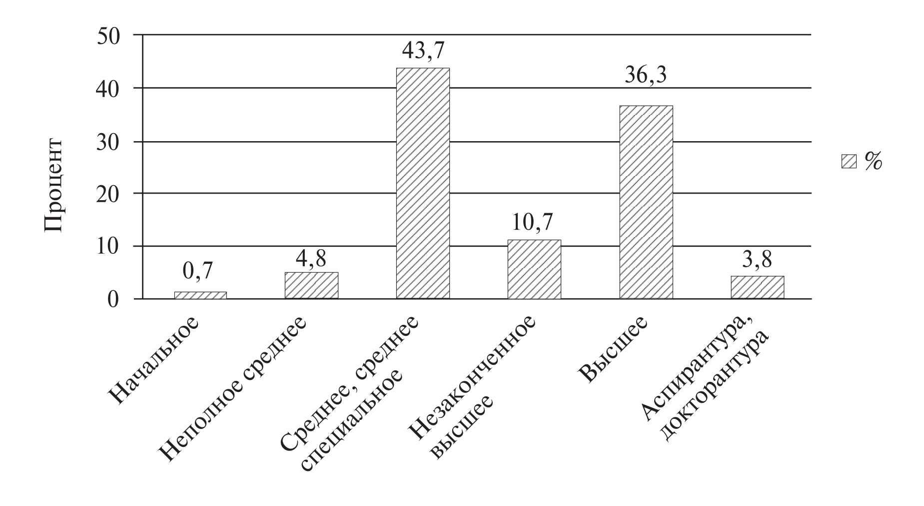


Рис. 3. Распределение выборочной совокупности по уровню образования

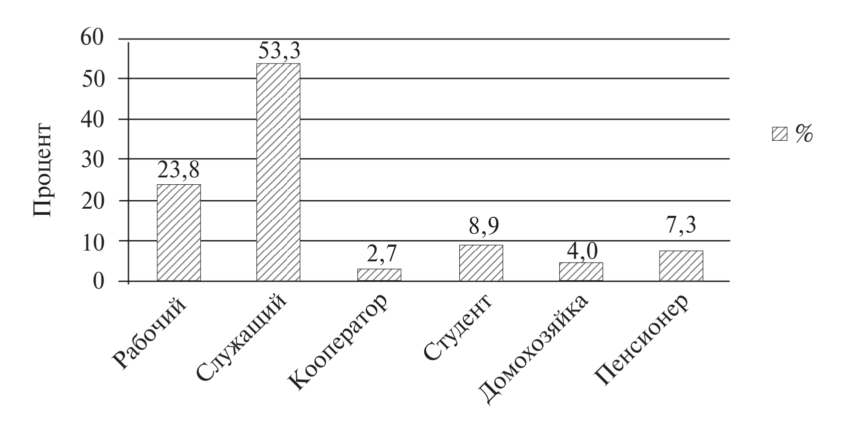


Рис. 4. Распределение выборочной совокупности по роду занятий

Определение факторной структуры восприятия избирателями личностного образа народного депутата

Стереотипы обыденного сознания могут быть охарактеризованы через смысловое пространство, описывающее семантические поля соответствующих ролевых позиций (персонажей). Для выявления стереотипов обыденного сознания относительно типологии личности (типажности) депутата мы использовали метод личностного семантического дифференциала (ЛСД).

Метод ЛСД выступает комбинацией метода контролируемых ассоциаций и процедур шкалирования. Процедура ЛСД может быть охарактеризована как вербальная модель поведенческого реагирования на персонажи с разным ролевым статусом. ЛСД ориентирован на разделение не персонажей, а реакций испытуемых на эти персонажи.

В целях экспериментальной психосемантической реконструкции стереотипов сознания избирателей в качестве стимулов в ЛСД были взяты персонажи «Самый распространенный тип народного депутата», «Тип народного депутата, который Вам нравится», «Тип народного депутата, который Вам не нравится», «Самый распространенный тип партийного руководителя», «Я сам (сама)». Выбор персонажей-стимулов обусловлен установленным в социально-психологических исследованиях фактом, что системообразующим элементом категориальной системы обыденного сознания при познании и восприятии социальных субъектов является «оценочное противопоставление». Оно выступает в роли «оси» нормативной оценки, на одном полюсе которой оказываются субъекты — носители социально одобряемых характеристик, а на другом — «негативные» субъекты.

В ходе пилотажного эксперимента к 80 испытуемым (в их роли выступили студенты МГУ им. М. В. Ломоносова) обратились с просьбой оценить восемь соответствующих исследованию персонажей («Депутат, чьи взгляды вызывают у меня возражение», «Депутат, олицетворяющий успех» и т. п.) по 60 униполярным шкалам, которые образованы сравнительно высокочастотными «личностными» прилагательными, имеющими минимальную оценочную нагрузку. При составлении исходного списка прилагательных учитывались также построенные на материале русского языка личностные семантические дифференциалы [8–10]. Полученные в результате суммарные групповые матрицы подверглись факторному анализу. В итоге отобрано 30 прилагательных, максимально коррелировавших с этими персонажами и образующих их противоположные полюса. Выделенные прилагательные скомплектованы в 14 биполярных шкал-дескрипторов – основу ЛСД (табл. 1).

Далее респондентам основного исследования предложили по данным шкалам оценить каждого из перечисленных в предыдущем абзаце персонажей.

Таблица 1

Дескрипторные шкалы персонажей			Тип депутата, который Вам нравится	Тип депутата, который Вам не нравится
Открытый	3 2 1 0 1 2 3	Скрытный	<= +	– =>
Расчетливый	3 2 1 0 1 2 3	Проницательный	=> –	+ <=
Неформальный	3 2 1 0 1 2 3	Формальный	<= +	– =>
Сухой	3 2 1 0 1 2 3	Общительный	=> –	+ <=
Простой	3 2 1 0 1 2 3	Утонченный		<=
Возбудимый	3 2 1 0 1 2 3	Спокойный	=> –	+ <=
Неповторимый	3 2 1 0 1 2 3	Типичный	<= +	– =>
Строгий	3 2 1 0 1 2 3	Терпимый	–	+ <=
Простодушный	3 2 1 0 1 2 3	Хитрый	+	– =>
Интуитивный	3 2 1 0 1 2 3	Логичный	=> –	+ <=
Сдержанный	3 2 1 0 1 2 3	Порывистый	<= +	–
Властный	3 2 1 0 1 2 3	Мягкий	–	+ <=
Уступчивый	3 2 1 0 1 2 3	Упрямый		
Надменный	3 2 1 0 1 2 3	Доброжелательный	=> –	+ <=

Оценки персонажей по отдельным шкалам коррелируют друг с другом; с помощью факторного анализа можно выявить «пучки» таких высококоррелирующих шкал и сгруппировать их в факторы.

С социологической точки зрения каждый фактор можно рассматривать как смысловой инвариант некоторого множества шкал (множества дескрипторов, входящих в «пучок» корреляций), по которым люди оценивают личностные свойства народного депутата. Факторы выступают здесь своего рода «метаязыком» описания признаков, служащих респондентам для классификации персонажей. Можно сказать, что факторные структуры ЛСД отражают присущие респондентам структуры категоризации, опосредующие восприятие политического деятеля или самого себя, обыденную «типологию личности», выработанную житейским опытом респондентов. Использовавшийся в исследовании метод ЛСД как раз и направлен на выявление «имплицитной теории личности депутата»[1], присущей обыденному сознанию.

В дальнейшем мы приведем для краткости факторные структуры лишь двух персонажей.

Для персонажа «тип народного депутата, который Вам нравится» методом альфа-факторного анализа выделены четыре фактора.

П е р в ы й (объясняет 20,5 % суммарной дисперсии) объединяет на положительном полюсе дескрипторы «надменный – доброжелательный», «сухой – общительный», «возбудимый – спокойный» и может быть проинтерпретирован как «*коммуникативные свойства*».

В т о р о й (объясняет 12,4 % суммарной дисперсии) включает шкалы «простодушный – хитрый», «простой – утонченный» и может быть назван «*простота*» («бесхитростность»).

Т р е т и й (объясняет 10,7 % суммарной дисперсии) объединяет шкалы «неповторимый – типичный», «неформальный – формальный» и может быть интерпретирован как «*оригинальность*» («необычность»).

Ч е т в е р т ы й (объясняет 8,9 % суммарной дисперсии) содержит дескрипторы «строгий – терпимый», «властный –мягкий» и может быть назван «*властность*».

Весьма похожие результаты для данного персонажа дало и многомерное неметрическое шкалирование по методу Джонсона, с помощью которого были выделены три фактора (рис. 5–7; для удобства рассмотрения дескрипторные пары описаны только первыми дескрипторами, а пло-

[1] Под «имплицитной теорией личности» понимается совокупность неявных представлений индивида о связях между личностными чертами народного депутата. На основании частичных, случайных впечатлений о личностных особенностях депутата имплицитная теория личности дает возможность сформировать его целостный образ.

скость, на которую они проецируются, задается одним и тем же фактором по обеим осям). В качестве меры различия использовалась величина s = (1 – *r*) / 2, где *r* – коэффициент корреляции между двумя векторами оценок персонажей по отдельным шкалам.

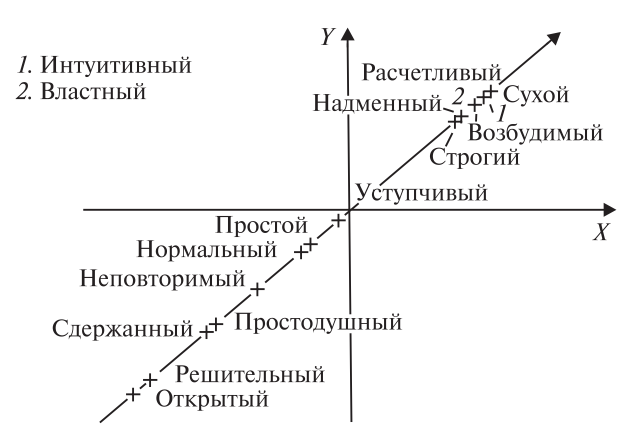


Рис. 5. Проекция конфигурации дескрипторов персонажа «тип депутата, который нравится» на ось, образованную фактором «коммуникативные свойства»

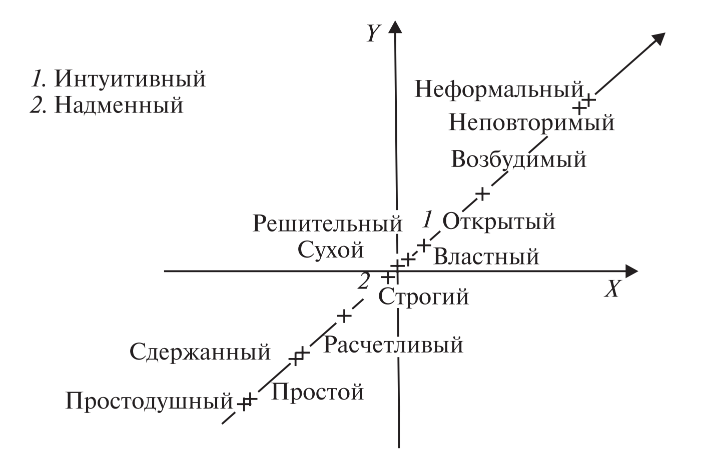


Рис. 6. Проекция конфигурации дескрипторов персонажа «тип депутата, который нравится» на ось, образованную фактором «простота – оригинальность»

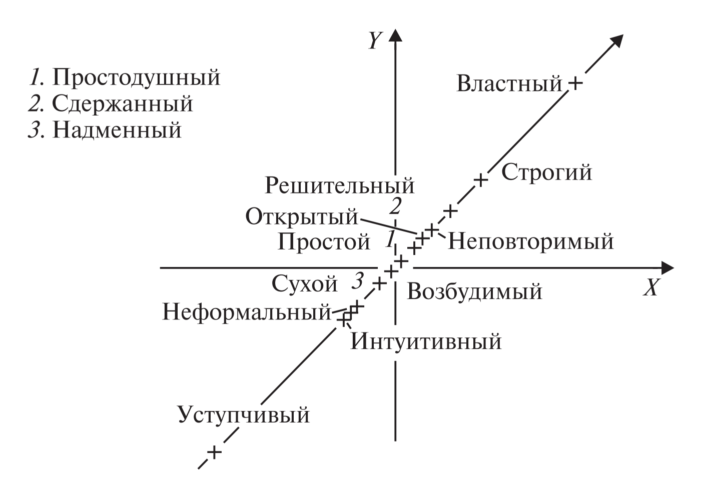


Рис. 7. Проекция конфигурации дескрипторов персонажа «тип депутата, который нравится» на ось, образованную фактором «властность»

Первый фактор здесь наиболее полно характеризуется дескрипторами «открытый — скрытный», «расчетливый — проницательный», «сухой — общительный», «возбудимый — спокойный», «интуитивный — логичный» и, вероятнее всего, также имеет смысл *коммуникативных свойств* персонажа, его «*стиля общения*», хотя здесь явно просматриваются и элементы «стиля *мышления*».

Второй фактор имеет на своих полюсах дескрипторы «простой — утонченный», «простодушный — хитрый», «неформальный — формальный», «неповторимый — типичный» и, очевидно, объединяет факторы № 2 и № 3, выявленные с помощью альфа-факторного анализа, т. е. факторы «*простота*» и «*оригинальность*».

Третий фактор характеризуется дескрипторами «уступчивый — упрямый», «властный — мягкий», «строгий — терпимый» и интерпретируется как «*властность*» или «*сила Я*».

Для персонажа «тип народного депутата, который Вам не нравится», методом альфа-факторного анализа также определены четыре фактора.

Первый (объясняет 19,8 % суммарной дисперсии) включает в себя переменные «надменный — доброжелательный», «сухой — общительный», «возбудимый — спокойный». Данный фактор может быть интерпретирован как «*коммуникативные свойства*» (естественно, что для отрицательного персонажа эти свойства имеют негативную окраску).

Второй (объясняет 12,1 % суммарной дисперсии) содержит дескрипторы «простодушный — хитрый» и «простой — утонченный». Этот набор переменных, знакомый нам по предыдущему персонажу, может быть условно назван «*простотой*».

Т р е т и й (объясняет 10,2 % суммарной дисперсии) объединяет переменные «властный — мягкий» и «строгий — терпимый» и может быть представлен как «*властность*».

Ч е т в е р т ы й (объясняет 9,1 % суммарной дисперсии) содержит дескрипторы «типичный — неповторимый» и «формальный — неформальный». Это уже известный нам по предыдущему персонажу фактор «*оригинальность*».

Итак, факторная структура персонажа «тип депутата, который Вам не нравится» очень похожа на факторную структуру персонажа «тип народного депутата, который Вам нравится». Единственное качественное различие заключается в том, что в структуре персонажа «тип народного депутата, который Вам не нравится», по сравнению с позитивным персонажем, переставлены местами 3-й и 4-й факторы. Таким образом, при восприятии данного персонажа фактор «*властность*» оказывается для респондентов более важным, чем «*оригинальность*».

Многомерное шкалирование по методу Джонсона дало для персонажа «тип народного депутата, который Вам не нравится» результаты, почти идентичные тем, что и для персонажа «тип депутата, который Вам нравится».

Таким образом, результаты многомерного неметрического шкалирования хорошо согласуются с выводами альфа-факторного анализа. Отсюда следует, что выделенные факторы, скорее всего, являются не артефакторами, а действительными «координатными осями» восприятия народного депутата.

Устойчивость данной факторной структуры косвенно проверялась далее в одном из избирательных округов г. Москвы уже в разгар предвыборной кампании (январь 1990 г.). Выборка представляла избирателей данного округа ($N = 300$). В рамках исследования определялась структура восприятия избирателями персонажа «Ваш идеал политика». Совокупность дескрипторов, описывающих личностные качества данного персонажа, несколько отличалась от ранее представленной (табл. 2). Тем не менее факторная структура персонажа оказалась полностью идентичной уже рассмотренным.

П е р в ы й ф а к т о р (объясняющий 16,2 % суммарной дисперсии) интегрирует дескрипторы «надменный — доброжелательный», «сухой — общительный» и известен нам как «*коммуникативные свойства*».

В т о р о й ф а к т о р (объясняющий 13,9 % суммарной дисперсии) образован дескрипторами «простодушный — хитрый», «простой — сложный» и может быть интерпретирован как «*простота*».

Т р е т и й ф а к т о р (объясняющий 10,0 % суммарной дисперсии) состоит из дескрипторов «неповторимый — типичный», «неформальный — формальный» и как в предыдущих случаях назван «*оригинальность*».

Ч е т в е р т ы й ф а к т о р (объясняющий 7,9 % суммарной дисперсии) сформирован дескриптором «властный – мягкий» и интерпретируется как «*властность*».

Таблица 2

Дескрипторные шкалы для персонажа «Ваш идеал политика»		
Открытый	3 2 1 0 1 2 3	Скрытный
Рациональный	3 2 1 0 1 2 3	Эмоциональный
Неформальный	3 2 1 0 1 2 3	Нормальный
Сухой	3 2 1 0 1 2 3	Общительный
Простой	3 2 1 0 1 2 3	Сложный
Возбудимый	3 2 1 0 1 2 3	Спокойный
Неповторимый	3 2 1 0 1 2 3	Типичный
Подвижный	3 2 1 0 1 2 3	Устойчивый
Простодушный	3 2 1 0 1 2 3	Хитрый
Обаятельный	3 2 1 0 1 2 3	Деловой
Умеренный	3 2 1 0 1 2 3	Настойчивый
Властный	3 2 1 0 1 2 3	Мягкий
Предсказуемый	3 2 1 0 1 2 3	Неожиданный
Надменный	3 2 1 0 1 2 3	Доброжелательный
Решительный	3 2 1 0 1 2 3	Осторожный

Оценка персонажей ЛСД в пространстве выделенных факторов

На практике, например, при проведении избирательной кампании конкретным кандидатом или прогнозировании возможных результатов выборов, важно не только выявить основные факторы, составляющие структуру восприятия личностного образа народного депутата, но и определить, как в пространстве этих факторов оцениваются вышеназванные персонажи.

Можно сказать, что факторы – это такие характеристики персонажей, которые существенны для принятия избирателем решения относительно кандидата в народные депутаты. Иными словами, социально-психологический смысл факторов состоит в подготовке субъектом решения относительно того или иного персонажа. Стереотипный характер отношения к кандидату заключается в том, что решение уже подготовлено отнесением его к сформированной заранее категории, оценкой по уже сложившейся системе факторов. Можно сказать, что «пространство восприятия» кан-

дидатов выступает прообразом «политических реакций» на них. Поэтому идеальному депутату важно вписаться в факторы восприятия «идеального политика» или «типа народного депутата, который нравится».

На основе данных опроса определены средние значения (медианы) оценок дескрипторов для обоих персонажей, описанных в предыдущем разделе, что в совокупности позволило определить полюса (положительный и отрицательный) дескрипторных пар, представленных в табл. 1. Когда средняя оценка дескриптора по персонажу сдвинута в ту или иную сторону, это фиксируется соответствующими стрелками «<=» и «=>». Отсутствие у дескрипторной пары стрелок означает, что среднее значение оценки равно 0, т. е. респонденты не смогли однозначно оценить какой-либо из признаков пары как позитивный или негативный для данного персонажа. Исходя из оценок дескрипторов для обоих персонажей можно охарактеризовать (оценить) каждый фактор, составляющий структуру восприятия респондентами данного персонажа.

Для персонажа «тип народного депутата, который Вам нравится» фактор «*коммуникативные свойства*» характеризуется положительными значениями входящих в него дескрипторов («доброжелательный», «общительный», «спокойный»). Фактор «*простота*» никак не оценен, и это означает, что данный персонаж воспринимается как в меру «простой» и в меру «хитрый» без преобладания одного из этих двух признаков. Оценки дескрипторов, входящих в фактор «*оригинальность*», сдвинуты в положительную сторону, и следовательно, персонаж «тип депутата, который нравится» в сознании большинства избирателей все-таки чем-то выделяется из общего окружения, обладает какой-то «неповторимой» чертой. Наконец, фактор «*властность*», так же как и простота, имеет нулевые оценки входящих в него дескрипторов, т. е. оцениваемый персонаж, по мнению респондентов, не слишком «строг» и не слишком «мягок».

Для персонажа «тип народного депутата, который Вам не нравится» фактор «*коммуникативные свойства*» имеет, естественно, отрицательную окраску («надменный», «сухой», «возбудимый»). Оценка фактора «*простота*» достаточно противоречива: отрицательный тип депутата и «прост», и «хитер» одновременно. В оценке фактора «*властность*» преобладают уже с негативной окраской признаки «строгий» и «властный» (вероятно, чересчур «властный» или слишком «строгий»). Наконец, фактор «*оригинальность*» становится для данного персонажа антиподом, т. е. «типичностью» и «формальностью».

Таким образом, стараясь походить по личностным качествам на «тип народного депутата, который нравится» и, наоборот, смягчая в своем политическом имидже (образе) качества, которые присущи «типу народного депутата, который не нравится», кандидат имеет больше шансов на успех в предвыборной борьбе.

Вместе с тем важно знать не только оценки персонажей ЛСД по каждому фактору, но и значимость (вес) этих факторов в структуре восприятия респондентом каждого персонажа. Нами проведено многомерное индивидуальное шкалирование персонажей ЛСД, в результате которого построена конфигурация персонажей в пространстве выделенных факторов (рис. 8–9, названия персонажей приведены в сокращенном виде).

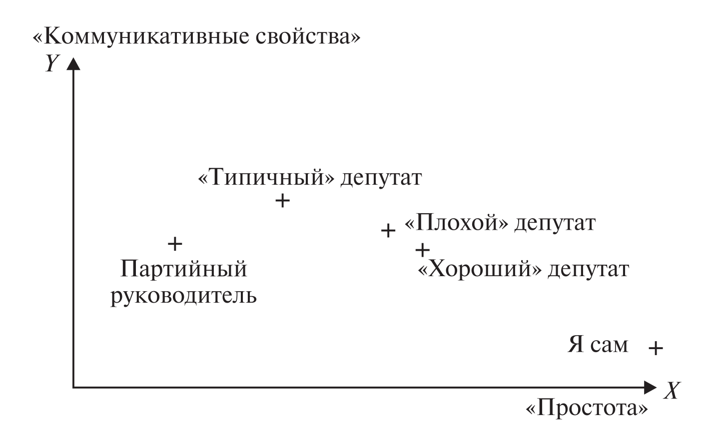


Рис. 8. Проекция конфигурации персонажа ЛСД на плоскость, образованную факторами «коммуникативные свойства» и «простота»

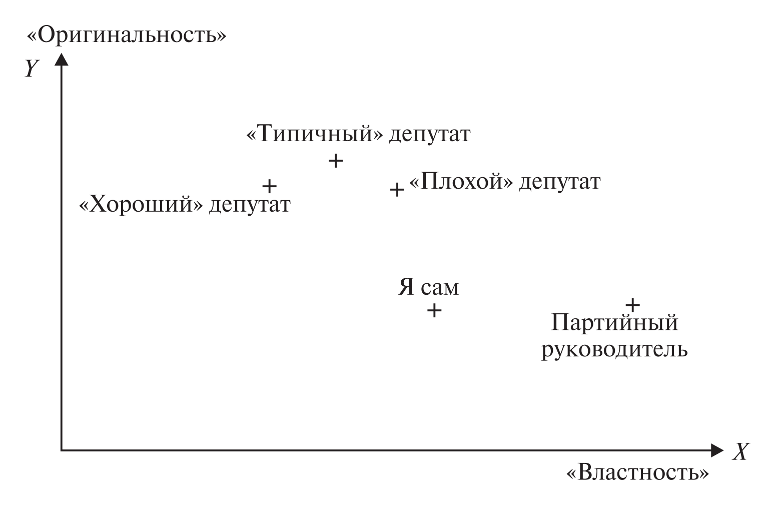


Рис. 9. Проекция конфигурации персонажа ЛСД на плоскость, образованную факторами «оригинальность » и «властность»

Из анализа конфигурации следует, что фактор «*коммуникативные свойства*» значимо входит в структуру восприятия всех персонажей, кроме персонажа «я сам»; фактор «*простота*» («бесхитростность»), напротив, более всего связан (коррелирует) с персонажем «я сам» и почти не свойствен персонажу «самый распространенный тип партийного руководителя»; фактор «*оригинальность*» имеет довольно большое значение при восприятии персонажей-«депутатов» и незначительное при восприятии персонажей «я сам» и «партийный руководитель»; наконец, фактор «*властность*» («сила Я») в наибольшей степени коррелирует с персонажем «партийный руководитель» и в наименьшей – с персонажем «тип депутата, который нравится».

Конфигурация персонажей ЛСД в 4-факторном пространстве наглядно показывает степень различия в восприятии респондентами отдельных персонажей. Можно сказать, что структура восприятия различных персонажей-«депутатов» почти одинакова для них и не зависит от типа воспринимаемого персонажа – положительного или отрицательного. Напротив, она сильно отличается от структур восприятия персонажей «партийный руководитель» и «я сам», которые в свою очередь разнятся между собой. Данный вывод подтверждается результатами иерархического кластерного анализа и построения размытой классификации по методу Заде (рис. 10).

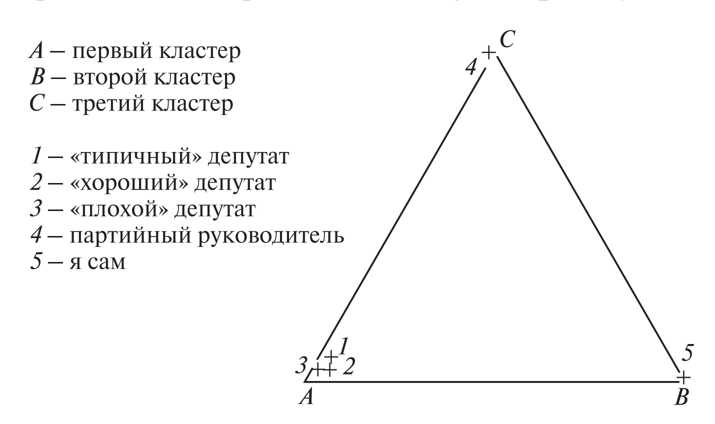


Рис. 10. Результаты размытой классификации персонажей ЛСД

Такое несовпадение структур восприятия можно объяснить тем, что персонаж «партийный руководитель», как и персонаж «я сам», в отличие от персонажей-«депутатов», по-видимому, не являются для избирателей политическими персонажами (публичными политиками) и воспринимаются совсем в другой категориальной сетке. Кроме того, персонаж «я сам» не является каким-либо социальным типом, а уникален для каждого респондента.

Другие примеры определения факторной структуры восприятия личностного образа народного депутата

При определении факторной структуры личностного образа очень важно соблюсти требование репрезентативности базового набора дескрипторов [11]. В конечном итоге исследователь всегда ограничивает пространство возможных реакций респондента, предлагая ему тот или иной выбор личностных качеств депутата, и следовательно, косвенно участвует в моделировании факторного пространства восприятия избирателем личностного образа депутата. Это отрицательное влияние на респондента можно уменьшить, либо увеличив число предлагаемых для оценки дескрипторов, либо предоставив респонденту право самому определить исходный набор шкал (личностных черт депутата). В рамках нашего исследования был проведен подобный эксперимент.

Испытуемым (300 человек – простая случайная выборка, репрезентирующая взрослое население одного из районов г. Москвы) предъявлялись сгруппированные в диады репертуарные позиции «депутат, который нравится Вам как личность», «депутат, который не нравится Вам как личность», «самый распространенный тип народного депутата», «депутат, олицетворяющий успех», «высоконравственный депутат», «депутат, чьи взгляды вызывают у Вас сильное раздражение», «депутат, чьи взгляды Вы разделяете». От испытуемого требовалось, подставив на место каждой репертуарной позиции конкретного народного депутата СССР (из предложенного списка наиболее известных депутатов), определить и назвать, в чем состоит существенное сходство или различие в диаде. Например, «депутат, олицетворяющий успех» (Горбачев) и «высоконравственный депутат» (Сахаров) – оба конструктивны.

Получив от испытуемых индивидуальные наборы шкал (субъективные категории), мы отобрали три наиболее часто встречающиеся категории. Ими оказались «критичность», «конструктивизм» и «жизненный путь».

К сожалению, пилотажный характер эксперимента не позволил с достаточной точностью определить более широкий набор дескрипторов, репрезентирующих личностные качества депутата. Тем не менее можно утверждать, что при восприятии (оценивании) уже известных избирателю депутатов наиболее существенными для него факторами являются, с одной стороны, «критичность», т. е. настроенность на вскрывание «всех и всяческих» недостатков и ошибок, с другой – «конструктивность», т. е. способность выдвигать (предлагать) реальные варианты решения проблем и исправления этих ошибок.

Некоторым образом такой вывод подтверждается результатами анализа отношений к различным личностным качествам народного депутата.

Респондентам основного исследования ($N = 1431$) предложили оценить значимость для них тех или иных личностных качеств депутата (табл. 3; оценка «3» означает, что данное качество, по мнению респондента, очень существенно для депутата, оценка «–3» – качество совершенно не существенно для депутата). Выделенные нами в тезаурусе личностных качеств народного депутата категории не могут служить научными категориями социологии или социальной психологии, здесь используются категории обыденного сознания, задающие структуру восприятия населением образа народного депутата.

Таблица 3

Личностные качества народного депутата	Используемая шкала	Значение медианы
Критичность	-3 -2 -1 0 1 2 3	3
Принципиальность	-3 -2 -1 0 1 2 3	3
Жизненный опыт	-3 -2 -1 0 1 2 3	3
Демократичность	-3 -2 -1 0 1 2 3	3
Личное обаяние	-3 -2 -1 0 1 2 3	2
Миролюбие	-3 -2 -1 0 1 2 3	2
Интеллигентность	-3 -2 -1 0 1 2 3	3
Самостоятельность	-3 -2 -1 0 1 2 3	3
Ответственность	-3 -2 -1 0 1 2 3	3
Профессиональные знания	-3 -2 -1 0 1 2 3	3
Простота	-3 -2 -1 0 1 2 3	1
Привлекательность	-3 -2 -1 0 1 2 3	1
Рациональность	-3 -2 -1 0 1 2 3	2
Темперамент	-3 -2 -1 0 1 2 3	1
Новаторство	-3 -2 -1 0 1 2 3	2

Данные опроса подверглись процедуре факторного анализа, в результате которого была построена факторная структура пространства личностных факторов.

Первый фактор (объясняет 28,6 % суммарной дисперсии) объединяет такие качества депутата, как «личное обаяние», «привлекательность».

Второй фактор (объясняет 13,0 % суммарной дисперсии) объединяет морально-нравственные качества: «критичность», «принципиальность», «ответственность».

Третий фактор (объясняет 7,3 % суммарной дисперсии) характеризуется профессионально-деловыми качествами народного депутата и объединяет «рациональность», «новаторство», «профессиональные знания».

Наиболее информативным (наиболее разделяющим депутатов при восприятии их избирателями) снова, как и при анализе факторной структуры персонажей ЛСД, оказался «внешний» фактор «личного обаяния». Однако при анализе средних оценок значимости (см. значение медианы оценок в табл. 3) выяснилось, что наиболее существенными респонденты считают морально-нравственные качества депутата, далее по важности следуют профессионально-деловые качества, и лишь затем внешняя «привлекательность».

Сравнивая результаты разных экспериментов, можно сделать следующие выводы:

1. В случае, когда конкретный народный депутат (или кандидат в депутаты) еще не известен избирателю, наиболее важными (информативными) для последнего являются «внешние» характеристики депутата, такие как «личное обаяние», «понятность» («простота»), определенная (небольшая) «оригинальность».

2. При восприятии депутатов (кандидатов), которые в той или иной степени уже известны респондентам, главную роль играют морально-нравственный фактор, объединяющий такие личностные качества, как «критичность», «принципиальность», «ответственность», «властность» (последнее чаще выступает как негативный признак), и профессионально-деловой фактор, включающий «конструктивность», «рациональность», «новаторство», «профессиональные знания». При этом значимость профессионально-деловых качеств все-таки несколько уступает значимости качеств морально-нравственных.

Заключение

Как показывает опыт мировой науки, поведение избирателей невозможно предсказать исключительно на основании знания наличной ситуации — его можно прогнозировать только на основании знания моделей социальной действительности, которые существуют в сознании людей. Для того чтобы предсказать, кто из кандидатов в народные депутаты будет популярен у населения, необходимо узнать порожденный массовым сознанием специфический «образ предпочтительного депутата», раскрыть содержание личностных стереотипов, определить ту систему значений, категориальную сетку, при помощи которой избиратели вычленяют в народном депутате, политике значимые для себя признаки.

Многомерность социально-психологического описания личности народного депутата сама по себе вряд ли вызывает сомнение. Спорным является скорее набор личностных черт, подлежащих описанию. Из множества личностных качеств нужно выделить некоторые интегральные социально-психологические характеристики. Выделенные нами методами альфа-факторного анализа и многомерного шкалирования и описанные выше фак-

торы отражают, как мы полагаем, основания классификации, которыми сознательно или неосознанно пользуются избиратели при сопоставлении личностей кандидатов в народные депутаты. Данные факторы являются, на наш взгляд, довольно устойчивыми элементами структуры личностного образа народного депутата, порожденного массовым сознанием избирателей.

Литература

1. *Демидов А. М.* Секреты избирателей // Социол. исслед. 1989. № 5.
2. *Бритвин В.Г.* Кто станет депутатом? // Социол. исслед. 1989. № 6.
3. *Комаровский В. С.* Типология избирателей // Социол. исслед. 1990. № 1.
4. *Агеев В. С.* Межгрупповое взаимодействие: социально-психологические проблемы. М. : МГУ, 1990.
5. *Ким Дж.-О., Мьюллер Ч. У.* Факторный анализ: статистические методы и практические вопросы // Факторный, дискриминантный и кластерный анализ / под ред. И. С. Енюкова. М., 1989.
6. *Терехина А. Ю.* Анализ данных методами многомерного шкалирования. М. : Наука, 1986.
7. *Качанов Ю. Л., Горбачев О. Г.* Метод построения представительной выборки по оценкам моментов второго порядка при ограничении на стоимость измерения. М. : ИС АН СССР, 1990.
8. *Бажин Е. Ф., Эдкинд А. М.* Личностный семантический дифференциал : метод. рек. Л. : ЛГУ, 1983.
9. *Петренко В. Ф.* Введение в экспериментальную психосемантику: исследование форм репрезентации в обыденном сознании. М. : МГУ, 1983.
10. *Шмелев А. Г.* Об устойчивости факторной структуры личностного семантического дифференциала // Вестн. МГУ. Сер. 14. Психология. 1982. № 2.
11. *Франселла Ф., Баннистер Д.* Новый метод исследования личности. М. : Прогресс, 1987.

БАЗОВЫЕ ПОЛИТИЧЕСКИЕ ЦЕННОСТИ НАСЕЛЕНИЯ РОССИИ[1]

Ю. Л. Качанов, Г. А. Сатаров

В начале июня 1991 года Центр прикладных политических исследований ИНДЕМ провел социологическое выборочное исследование, направленное на изучение базовых политических ценностей населения Российской Федерации. Выборка репрезентировала взрослое население Российской Федерации в возрасте от 18 до 60 лет по полу, возрасту и образованию. Исследованием было охвачено 18 населенных пунктов – 12 го-

[1] Опубликовано: Информационные материалы. М. : Центр приклад. полит. исслед. ИНДЕМ, 1991.

родов и 6 сельских поселений. Опрос проводился в форме личного интервью по месту жительства. На анкету ответили 545 человек. Максимальная ошибка выборки при доверительном уровне 0,95 составила 9 %.

Анкета содержала 22 суждения, каждое из которых выражало одну из основных политических или экономических ценностей. Респонденты должны были выбрать из этого списка шесть суждений, наиболее точно отражающих их взгляды.

Предполагалось, что в обществе распространены определенные типы политического сознания (либеральный, люмпенский, фундаменталистско-коммунистический и т. п.) и сторонники того или иного типа выбирают определенные (возможно, несколько различающиеся) наборы предпочитаемых суждений, за которыми стоят те или иные политические ценности. Список суждений составлялся таким образом, чтобы их возможные комбинации могли представлять разнообразие политических позиций, описываемое двумя парами оппозиций: «политическая свобода – тоталитаризм» и «свободный рынок – плановое хозяйство» (таблица). Исследование было нацелено на выделение реально существующих в обществе типов политического мышления, описанных в терминах этих двух переменных, и изучение их распространенности.

№	Суждения анкеты
1	Если долго и усердно работать, то можно добиться успеха в жизни
2	Чтобы наша жизнь стала лучше, нужно укреплять государство
3	Надо предоставить народу больше прав при принятии важных государственных решений
4	В нашей стране есть много людей, которые мешают нам жить
5	Землю надо раздать желающим в частную собственность
6	Конкуренция – это хорошо, она побуждает людей работать напряженно и развивать новые идеи
7	Сейчас главное – сберечь социализм
8	Государство должно заботиться, чтобы у всех все было
9	Коммунизм – это то, ради чего стоит жить и работать
10	Богатеть можно только за счет других
11	Никакая экономия не может нормально работать без планирования
12	Надо работать, чтобы моя жизнь становилась лучше и чтобы было что оставить детям
13	Стране необходимы могучие вооруженные силы
14	Если в стране что-то разлаживается, надо вводить чрезвычайное положение
15	Государство должно заниматься своим делом и уважать мои права

№	Суждения анкеты
16	Если правильно все планировать и строго выполнять план, то в экономике все будет в порядке
17	Нельзя допустить, чтобы у нас люди еще больше различались по доходам и положению
18	Главное – это порядок и дисциплина
19	И в промышленности нужна настоящая частная собственность
20	Человек и государство должны быть равны перед законом
21	Безработные должны браться за любую имеющуюся работу или же терять пособие по безработице
22	Не важно, какой у нас будет строй, лишь бы жить по-человечески

На рис. 1 приведена диаграмма, изображающая дерево группировок суждений в представлениях респондентов. Суждения попадают в одну группу на диаграмме, если они часто выбираются респондентами в одну шестерку. Диаграмма построена методом кластерного анализа.

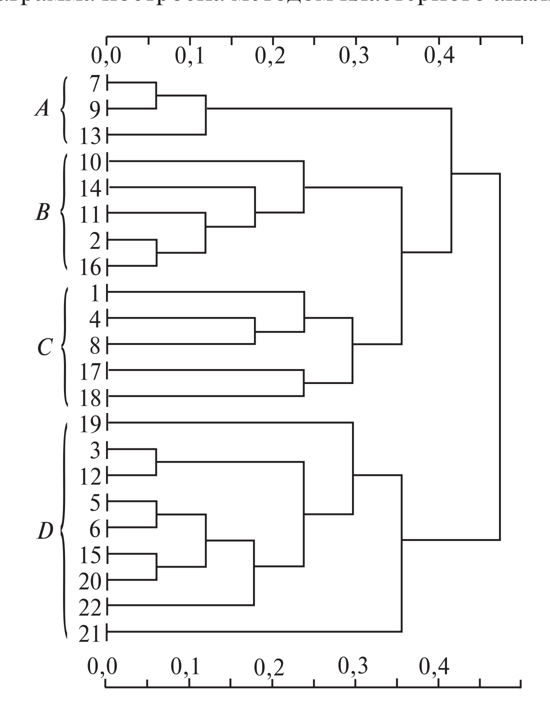


Рис. 1. Дерево группировки политических ценностей, построенное на основании выбора, осуществляемого респондентами

На диаграмме легко выделяются четыре основные группы суждений. Исходя из суждений, попавших в каждую группу, мы интерпретируем их следующим образом: группа *A* – коммунистическая группа (суждения 7, 9, 13), группа *B* – тоталитарная группа (10, 14, 11, 2, 16), группа *C* – социально-консервативная группа (1, 4, 8, 17, 18), группа *D* – либеральная группа (19, 3, 12, 5, 6, 15, 20, 22, 21). Поскольку эти группировки получены по результатам статистического анализа выборов суждений респондентами, постольку можно утверждать, что в обществе сложились четыре основных типа политического сознания, выражающиеся в предпочтениях соответствующих групп ценностей.

Теперь необходимо установить, как эти типы соотносятся с двумя характеристиками политического сознания, заложенными в схему эксперимента. Мы имеем в виду указанные выше оппозиции: «политическая свобода – тоталитаризм» и «свободный рынок – плановое хозяйство». Специальные статистические методы (например, многомерное шкалирование или факторный анализ) позволяют погрузить суждения в числовое пространство возможно меньшей размерности так, чтобы ассоциированные в представлениях респондентов суждения располагались по соседству, а сильно различающиеся по политическому смыслу лежали далеко друг от друга. Одновременно можно установить и число осей этого пространства.

Вычисления показали, что для адекватного отображения сходства/различия между суждениями достаточно одной оси. Эта ось изображена на рис. 2, как ось *X* (в условных единицах от 0 до 17). Суждения представлены на оси *X* столбиками, номера которых соответствуют номерам суждений. Высота каждого столбика отражает количество респондентов (в процентах), выбравших соответствующее суждение.

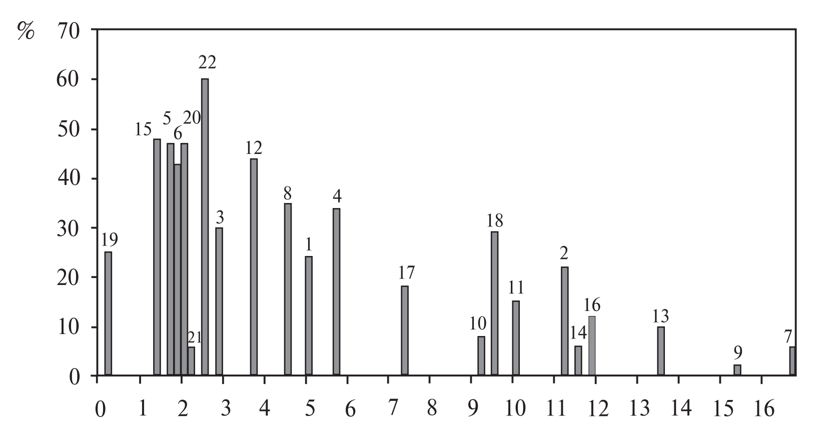


Рис. 2. Спектр предпочтений политических ценностей. Ось *X*: положительное направление – «тоталитаризм», отрицательное – «либерализм». Ось *Y* – доля респондентов (в % от числа опрошенных), выбравших каждое из 22 суждений

Расположение суждений на оси *X* позволяет интерпретировать ее следующим образом: положительное направление – «тоталитаризм», отрицательное направление – «либерализм». Легко видеть, что группировка суждений на рис. 1 соответствует их расположению на оси *X*.

Одномерность полученного решения означает, что в политическом сознании респондентов сливаются ценности политической и экономической свобод. На рис. 2 видно также, что в сознании респондентов ценности рыночной экономики противопоставлены традиционным коммунистическим ценностям. На этом основании можно предположить, что в глазах респондентов традиционные ценности «социалистического выбора» несовместимы с переходом к рыночной экономике.

На рис. 3 приведена гистограмма распределения респондентов на той же оси «либерализм – тоталитаризм». Координаты респондентов определялись как среднее координат выбранных ими суждений. Общий центр тяжести распределения смещен в сторону либеральных ценностей. Два основных пика в левой части гистограммы соответствуют двум ведущим типам политического сознания.

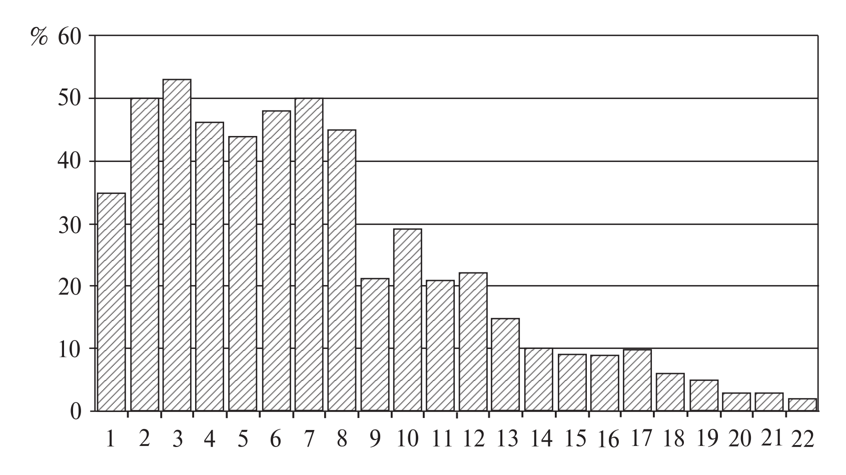


Рис. 3. Гистограмма распределения респондентов по оси «либерализм – тоталитаризм». Ось *X* – номера интервалов. Ось *Y* – количество респондентов, попавших в интервал

Важно установить распространенность установленных типов политического сознания среди респондентов. Нам удалось сделать это, используя один из методов построения так называемых размытых классификаций. Результаты вычислений приведены на рис. 4. Мы не считаем неожиданным, что сторонники ценностей группы *A* оказались в явном меньшинстве. Важно другое: имеет место хрупкий баланс между сторонниками либеральных и социально-консервативных ценностей: 47 % и

43 % соответственно. Из этого вытекают, как минимум, два следствия: первое – за демократических лидеров сейчас отдают голоса не только носители демократических убеждений, и второе – процесс демократических преобразований может легко быть заторможен, если огромная социально-консервативная масса качнется вправо. Следует учитывать, что социально-консервативное сознание поддерживает демократизацию ситуативно – из-за общего кризиса тоталитарной системы. Между тем социально-консервативные и тоталитарные ценности имеют общий корень. Поэтому в случае кризисной ситуации эти люди легко вернутся на позиции коммунистического фундаментализма.

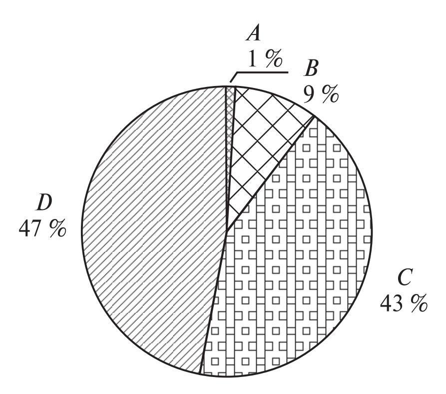


Рис. 4. Численность групп респондентов (в % от числа ответивших), поддерживающих коммунистическую группу ценностей (*A*), тоталитарную группу ценностей (*B*), социально-консервативную группу ценностей (*C*), либеральную группу ценностей (*D*). Данные получены методом размытой классификации

РАСЧЕТ РЕЙТИНГОВ ЗАКОНОДАТЕЛЕЙ (КОНСЕРВАТИЗМ И РАДИКАЛИЗМ НА II СЪЕЗДЕ НАРОДНЫХ ДЕПУТАТОВ СССР)[1]

Г. А. Сатаров, С. Б. Станкевич

Обыденное сознание, как и научная методология, создавая обобщенную картину мира, невольно упрощает, схематизирует ее. Если это мир политических деятелей, то обычно употребляется классификация (или шкалы) типа «левые – правые», «консерваторы – радикалы» и т. п. Однако очевидно, что содержания понятий «левые – правые» времен Великой

[1] Опубликовано: Демократические институты в СССР: проблемы и методы исследования. С. 6–17.

французской революции, военного коммунизма и перестройки отнюдь не тождественны, что консерваторы и радикалы в США и в СССР объединены или альтернативны на основе совершенно различных критериев. Более того, далеко не всегда многообразие политических позиций и политического поведения может быть адекватно описано единственной одномерной характеристикой, как бы она ни называлась – «консерватизм – радикализм» или иначе [1–3]. Ниже, для терминологической определенности мы используем оппозицию «консерватизм – радикализм».

Мысль о неоднозначности конкретного содержания понятий «консерватизм – радикализм» кажется банальной. Тем более удивительно, что эти понятия легко прививаются на самой разнообразной политической и культурной почве, легко и достаточно точно усваиваются общественным сознанием, не спешащим заменить их более конкретными. Значит, нечто универсальное, объективно инвариантное, делает эти понятия ясными, удобными и работающими.

В данной статье <...> описывается методика, с помощью которой универсальным категориям приписывается конкретное политическое содержание, порожденное конкретной исторической ситуацией. <...> В качестве объекта взят II Съезд народных депутатов СССР.

Анализ поименных голосований

При изучении законодательного органа возникают две основные проблемы: 1) что должно служить исходными данными; 2) в терминах каких переменных следует описывать субъекты законодательной деятельности. Наш опыт говорит: наиболее удобная информация – результаты поименных голосований. В них выражаются недвусмысленные и ответственные политические решения, регулярно принимаемые законодателями; поименные голосования образуют несвернутые данные в простом формализованном виде, позволяющем сопоставлять и законопроекты, и законодателей; содержание поименных голосований отражает наиболее насущные проблемы, встающие перед обществом.

Совокупность результатов поименных голосований образует простое описание законодателей точками многомерного пространства, размерность которого совпадает с числом голосований. Однако такое описание наследует все «шумы», содержащиеся в исходных данных, и, как правило, имеет слишком высокую размерность, что неудобно при анализе и моделировании. Опыт авторов [1, 2] и других исследователей [7] показывает, что известный тезис Альберта Эйнштейна «Бог изощрен, но не злонамерен» иногда оправдывается и в социальных науках. Анализируя данные о поименных голосованиях в Сенате США, мы установили, что разнообразие политических позиций сенаторов, выраженное разнообразными же исходами голосований, сводится к одной или двум числовым перемен-

ным, значения которых приписываются сенаторам. Эти переменные интерпретируются как идейно-политические характеристики политического сознания законодателей.

Приступая к анализу поименных голосований народных депутатов на II Съезде, мы полагали, что ни отсутствие политического опыта у законодателей (что может привести к несформированности или нестабильности критериев решений), ни аморфность и мозаичность структуры законодательного корпуса не должны повлиять на возможность существенного понижения размерности при переходе от поименных голосований к латентным идейно-политическим переменным.

Для проверки тезиса использовались данные, образованные всеми 24 поименными голосованиями на II Съезде. Напомним, что в них выделяются три основных сюжета: резолюции по докладу правительства, поправки к Конституции и резолюции по докладам комиссий съезда. В связи со вторым дебатировалась повестка дня, особенно вопрос о включении в нее 6-й статьи Конституции.

Для анализа поименных голосований применялась методика, описанная авторами в статье [1]. Было сформировано несколько выборок по 30 человек в каждой. При этом учитывалось следующее: отбирались депутаты с четкой политической позицией: они должны были принять или отвергнуть большую часть голосований, совокупность депутатов репрезентативно (по априорным оценкам) представляла политический спектр. Необходимость формирования выборок определялась главным образом техническими соображениями: методы углубленного анализа поименных голосований, позволяющие сопоставлять друг с другом все политические позиции, весьма сложны, и вычисления по более чем 2 тыс. депутатов одновременно практически не реализуемы. Однако результаты анализа выборок могут потом распространяться на всех законодателей, что показано ниже. В статье приведены результаты анализа для двух списков депутатов.

Последующий анализ основан на таком допущении: разумно полагать, что сходные (различные) политические позиции депутатов приводят к сходным (соответственно, различным) результатам поименных голосований у них (если, конечно, при голосовании они руководствовались своими политическими убеждениями, а не другими соображениями). Политическая позиция – это скрытая от исследователя информация, подлежащая выявлению, а результаты голосований – доступная информация, явно заданная и независимая от исследователя. Поэтому необходимо опираться на утверждение, обратное приведенному выше: сходным (различным) результатам поименных голосований двух депутатов соответствуют сходные (различные) политические позиции. Именно этим оправдан наш подход.

Для каждого списка депутатов вычислялась матрица различий между векторами голосований по всем парам депутатов. В качестве меры различия использовалась величина $S = (1 - r) / 2$, где r – коэффициент кор-

реляции между двумя бинарными векторами исходов голосований «да» и «нет» после исключения из обоих векторов тех голосований, по которым отсутствуют такие исходы хотя бы у одного депутата. Матрицы различий анализировались методами автоматической классификации и многомерного шкалирования [4, 6]. На рис. 1 приведены результаты кластерного анализа для первого списка депутатов. Использован метод построения аддитивных деревьев Саттаха и Тверски [8].

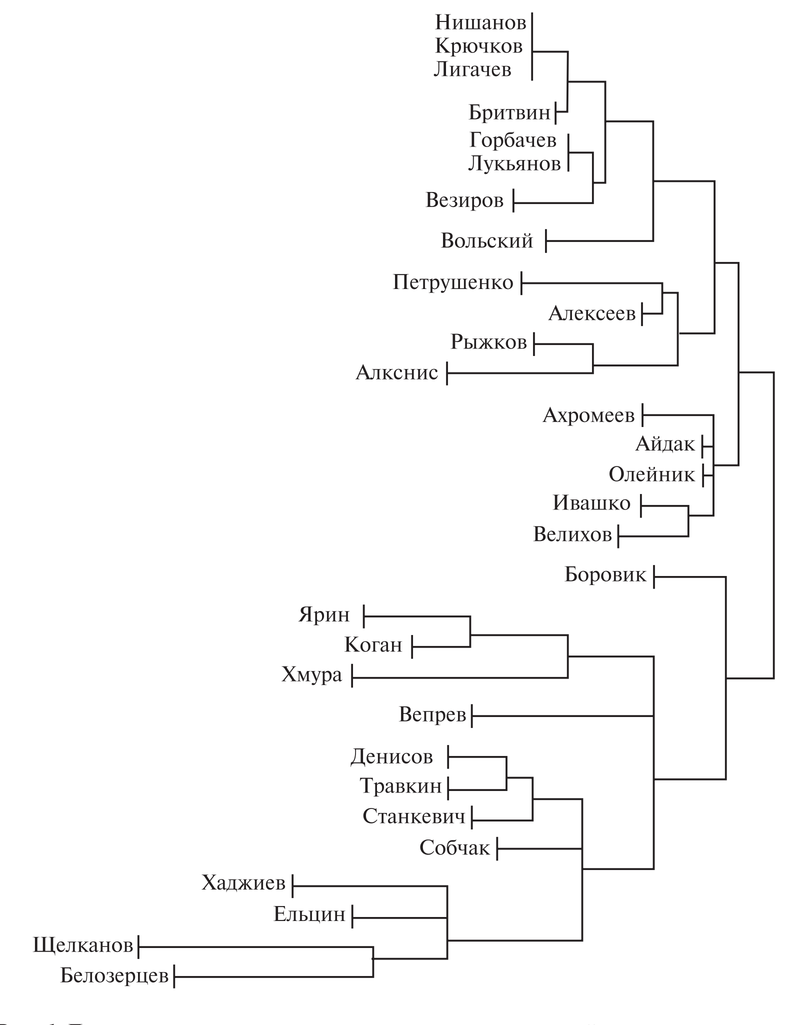


Рис. 1. Дендрограмма, полученная по матрице различий для первого списка депутатов методом аддитивных деревьев

Из дендрограммы видно, что группа депутатов разбивается на два кластера: «от Велихова и выше» и «от Боровика и ниже». Предварительный их анализ и сопоставление позволяют выдвинуть гипотезу: основной фактор, разделяющий депутатов на два кластера, обозначается как «поддержка/оппозиция действующему руководству» (далее эта гипотеза проверяется другими методами). Такая содержательная интерпретация дает возможность идентифицировать первый кластер как консервативный, второй – как радикальный. При более подробном анализе выделяются следующие кластеры: мощный и сплоченный кластер крайних консерваторов в составе – Нишанов, Крючков, Лигачев, Бритвин, Горбачев, Лукьянов, Везиров и стоящий несколько особняком Вольский; менее сплоченный, умеренно консервативный кластер в составе – Петрушенко, Алексеев, Рыжков, Алкснис; кластер консервативного центра – Ахромеев, Айдак, Олейник, Ивашко, Велихов. Радикальное крыло группы содержит кластеры: радикально-популистский – Ярин, Коган, Хмура; радикального центра – Денисов, Травкин, Станкевич, Собчак; крайних радикалов – Хаджиев, Ельцин, Щелканов, Белозерцев; несколько особняком стоят Боровик и Вепрев.

К интересным особенностям выявленной структуры отнесем сплоченные центристские группы, что противоречит общепринятой точке зрения на политическую ситуацию в стране, характеризующуюся как крайне поляризованная. Бросается в глаза (и подтверждается другими методами), что консервативное крыло существенно сплоченнее радикального. Это общая закономерность, которую авторы наблюдали в американском Сенате [3] и которая легко объяснима с точки зрения обсуждавшегося выше функционального разделения консерватизма и радикализма.

Дальнейший анализ осуществлялся методом неметрического многомерного шкалирования [4]. Прежде всего использовалась процедура вычисления оценки максимального правдоподобия размерности решения задачи многомерного шкалирования, описанной в [4]. С удивительным постоянством для каждой матрицы получалось значение 3. Однако анализ трехмерных решений показал, что третья ось образовывалась как результат совокупного действия малозначимых факторов, неучтенных первыми двумя осями. Поэтому в качестве основных рассматривались двумерные решения, ориентированные по главным осям облака рассеяния. Применен алгоритм неметрического многомерного шкалирования с коэффициентом соответствия Краскала второго типа.

На рис. 2 показано одно из плоских решений для приведенного выше списка депутатов. Полученный коэффициент соответствия равен 0,155, отношение дисперсий проекций на оси X и Y составляет 2,77.

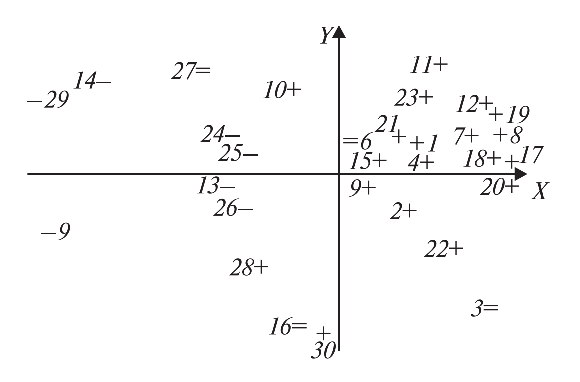


Рис. 2. Плоское решение задачи неметрического многомерного шкалирования для матрицы различий между результатами поименных голосований следующих депутатов:
1 – Айдак, *2* – Алексеев, *3* – Алкснис, *4* – Ахромеев, *5* – Белозерцев, *6* – Боровик, *7* – Бритвин, *8* – Везиров, *9* – Велихов, *10* – Вепрев, *1* – Вольский, *12* – Горбачев, *13* – Денисов, *14* – Ельцин, *15* – Ивашко, *16* – Коган, *17* – Крючков, *18* – Лигачев, *19* – Лукьянов, *20* – Нишанов, *21* – Олейник, *22* – Петрушенко, *23* – Рыжков, *24* – Собчак, *25* – Станкевич, *26* – Травкин, *27* – Хаджиев, *28* – Хмура, *29* – Щелканов, *30* – Ярин.
Знаками + («за»),– («против») и = («воздержался» или отсутствовал) – отмечены результаты голосования по постановлению об одобрении программы правительства. Начало координат совпадает с центром тяжести конфигурации

Для интерпретации осей использовались два подхода. Первый, неформальный, основан на сопоставлении депутатов, составляющих взаимную оппозицию в проекциях на каждую из осей, с учетом политических заявлений и действий. Второй, формальный, базируется на следующем тезисе: при адекватном решении задачи шкалирования для каждого голосования области точек, соответствующих депутатам, принявшим и отвергнувшим законопроект, должны эффективно линейно разделяться. (Для голосования по постановлению Съезда об одобрении программы правительства это проиллюстрировано на рис. 2.) Качество разделения характеризуется коэффициентом бисериальной корреляции [5] между проекциями на разделяющий вектор и исходами голосований. Если вектор имеет единичную длину, то его компоненты могут интерпретироваться как веса, характеризующие вклад каждой оси (и соответственно приписанного ей латентного идейно-политического фактора) в объяснение раскола по данному голосованию. Ниже приведена таблица, содержащая для некоторых голосований, помимо коэффициента бисериальной корреляции R и компонет разделяющего вектора Z_x и Z_y, вероятность ошибки P при отвержении гипотезы о случайном характере расположения точек относительно исхода данного голосования.

Соответствие решения задачи шкалирования результатам отдельных поименных голосований

Законопроект	*R*	*P*	Z_x	Z_y
Включение в повестку дня вопроса о Комитете Конституционного надзора	0,591	0,003	0,639	-0,769
Включение в повестку дня вопроса о 6-й статье Конституции	0,739	0,000	-0,998	0,055
Поддержка программы правительства	0,800	0,000	0,998	0,063
Открытие прений по докладу Комиссии о деятельности следственной группы Гдляна и Иванова	0,542	0,009	-0,111	0,994
Одобрение проекта постановления о советско-германском договоре о ненападении 1939 г.	0,664	0,000	0,108	0,994

Оси абсцисс соответствует характеристика, которой можно приписать оппозицию вида «за/против действующего руководства» (поддержка адекватна положительному направлению оси). Сложнее с осью ординат. Предположим, что ей соответствует характеристика, выражаемая оппозицией вида «динамика/консервация союзного устройства» (динамика соответствует положительному направлению оси). Решение хорошо согласуется с результатами кластерного анализа (см. рис. 2). Исключение составляет точка, соответствующая позиции депутата Рыжкова. Следует помнить, что решение задачи многомерного шкалирования может содержать неточности, порождаемые «шумами» в исходных данных. Однако неточности в координатах отдельных точек не очень существенны, поскольку объектом изучения являются факторы политического размежевания и наиболее значимые особенности политической структуры законодательного органа.

Для подтверждения устойчивости выделяемых факторов подобные исследования проводились и на других списках. Рис. 3 показывает результаты шкалирования для списка, частично пересекающегося с первым. Значение коэффициента соответствия для этого плоского решения составляет 0,150, отношение дисперсий проекций равно 3,26.

Депутаты, входящие в оба списка, занимают сходные места, и оси координат имеют ту же смысловую нагрузку. Тем самым полученная интерпретация факторов, определяющих раскол на II Съезде и конкретизирующих понятие «консерватизм – радикализм» подтверждается. Отметим, что в данном случае имеются два фактора, характеризующих это явление.

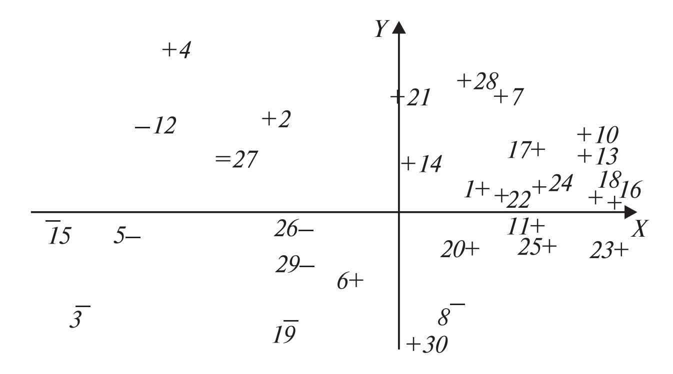


Рис. 3. Плоское решение задачи неметрического многомерного шкалирования для матрицы различий между результатами поименных голосований следующих депутатов:
1 – Айдак, *2* – Амонашвили, *3* – Белозерцев, *4* – Буачидзе, *5* – Бурулис, *6* – Вепрев, *7* – Вольский, *8* – Гидаспов, *9* – Гольданский, *10* – Горбачев, *11* – Дзасохов, *12* – Ельцин, *13* – Етылен, *14* – Заславская, *15* – Казанник, *16* – Крючков, *17* – Кугультинов, *18* – Лигачев, *19* – Лопатин, *20* – Месяц, *21* – Оганесян, *22* – Патон, *23* – Полозков, *24* – Разумовский, *25* – Скиба, *26* – Станкевич, *27* – Хаджиев, *28* – Шинкарук, *29* – Яворивский, *30* – Ярин.
Знаками + («за»), – («против») и = («воздержался» или отсутствовал) – отмечены результаты голосования по постановлению об одобрении программы правительства. Начало координат совпадает с центром тяжести конфигурации

Моделирование «идеального консерватора» и построение рейтингов депутатов

Полученные модели идейно-политического раскола позволяют искусственно сконструировать результаты поименных голосований депутата с любой наперед заданной позицией, представленной точкой в построенном двумерном пространстве. Проще всего это сделать, если депутат занимает какую-либо экстремальную позицию. Рассмотрим, например, «идеального консерватора», занимающего экстремальную позицию по оси *X* – ультраконсервативного сторонника действующей власти. Формальный критерий определения результатов его голосований предельно прост: он – «за» по любому голосованию, имеющему положительное значение первой компоненты вектора *Z*. Мы построили две последовательности поименных голосований – указанного «идеального консерватора» и его антипода «идеального радикала», у которого вектор поименных голосований противоположен вектору консерватора.

Для проверки построенных гипотетических моделей-депутатов из первого списка удалены двое (Айдак и Ахромеев), вместо них вставлены два

«экстремиста». Результаты шкалирования нового списка приведены на рис. 4. Значение коэффициента соответствия равно 0,140, отношение дисперсий проекций составляет 3,85. Места, занятые в конфигурации «идеальными» депутатами, указывают на правильность формирования «эталонных» депутатов.

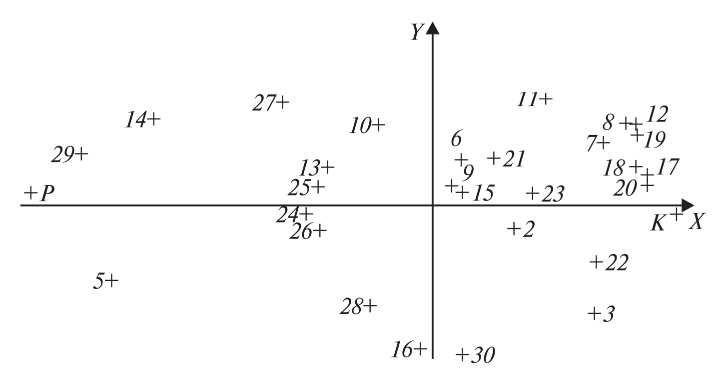


Рис. 4. Плоское решение задачи неметрического шкалирования для матрицы различий между результатами поименных голосований первого списка народных депутатов (без депутатов Айдака и Ахромеева). *K* – «идеальный консерватор», *P* – «идеальный радикал»

Эталоны дают возможность построения рейтингов для всех депутатов. Рейтинг строится как мера совпадения вектора поименных голосований некоторого депутата с эталонным вектором. Например, если в качестве эталона взят «идеальный консерватор» (см. рис. 4), то получится рейтинг консерватизма (при этом последний понимается как степень поддержки действующих властных структур).

Приведен простейший способ расчета рейтинга. Пусть результаты голосования некоторого депутата заданы вектором ($\alpha_1, \alpha_2 ... \alpha_n$), n – число голосований, причем $\alpha_i = 1$, если депутат проголосовал «за», и $\alpha_i = 0$ в противном случае; вектором ($\gamma_1, \gamma_2 ... \gamma_n$) характеризуется участие депутата в голосовании: $\gamma_i = 1$, если депутат проголосовал «за» или «против», $\gamma_i = 0$ в противном случае (т. е. он воздержался или не участвовал в голосовании); наконец пусть ($\varepsilon_1, \varepsilon_2 ... \varepsilon_n$) – вектор голосований «идеального консерватора». Вводим величину ν формулой

$$v = \frac{1}{n}\sum_{i=1}^{n}\gamma_i(2\alpha_i - 1)(2\varepsilon_i - 1)\,. \qquad (1)$$

Значение $v = 1$, если вектор депутата совпадает с эталонным вектором; $v = -1$ в противоположном случае; v тем ближе к 1, чем больше доля (относительного общего числа голосований n) случаев, когда депутат голо-

совал «за» или «против» одинаково с «идеальным консерватором». Формулу (1) можно модифицировать с тем, чтобы учесть обстоятельство, обсуждавшееся в предыдущем разделе: разные голосования в большей или меньшей степени связаны с первым фактором политического размежевания. Для этого вводим веса «важности» голосований. В данной работе использован чисто формальный подход, согласно которому в качестве весов брались абсолютные значения первых компонент векторов Z для каждого голосования. Пусть (w_1, w_2 ... w_n) – вектор весов. Тогда новый рейтинг определяется формулой

$$r = \frac{\sum_{i=1}^{n} w_i \gamma_i (2\alpha_i - 1)(2\varepsilon_i - 1)}{\sum_{i=1}^{n} w_i}. \quad (2)$$

Величина r, как и v, меняется в пределах от –1 до +1.

Однако часто для удобства рейтинг преобразуют так, чтобы он менялся от 0 (для «идеального радикала») до 100 (для «идеального консерватора»). Этого легко добиться, введя величину R, получающуюся из r (или v) преобразованием

$$R = \frac{r+1}{2} 100. \quad (3)$$

Ниже использованы рейтинги вида (3), полученные из величин (2). По этим первыми и подсчитывались средние значения и выборочные дисперсии для всевозможных групп депутатов, а также изучалось распределение рейтинга. На рис. 5 приведена гистограмма распределения рейтинга консерватизма для всех депутатов, голосовавших на II Съезде.

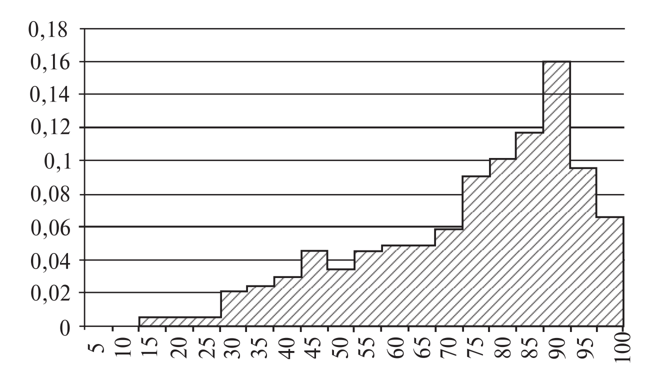


Рис. 5. Гистограмма распределения рейтингов депутатов по фактору поддержки действующих властных структур

Двумодальность распределения подтверждает наличие двух отчетливо выраженных группировок – консервативной и радикальной. Отождествление двух мод с такими группировками подтверждается, в свою очередь, тем, что первая мода совпадает со средним рейтингом Межрегиональной депутатской группы (напомним, что на II Съезде МДГ объявила о своей оппозиции; средние значения подсчитаны только по тем депутатам, которые подписали соответствующее заявление). Вторая (правая) мода очень близка к средним значениям рейтингов консервативности таких групп, как «руководство партии и правительства» и группа депутатов, избранных от Всесоюзной организации ветеранов войны и труда. Распределение средних рейтингов указанных групп еще раз свидетельствует о правильности интерпретации основного фактора политического размежевания на II Съезде народных депутатов СССР как фактора поддержки властных структур.

Литература

1. *Сатаров Г. А., Станкевич С. Б.* Голосования в Конгрессе США: опыт многомерного анализа // Социол. исслед. 1983. № 1.
2. *Сатаров Г. А., Станкевич С. Б.* Идеологическое размежевание в Конгрессе США // Социол. исслед. 1988. № 2.
3. *Станкевич С. Б., Сатаров Г. А.* Типологический анализ результатов голосований в Конгрессе США // Математические методы и ЭВМ в историко-типологических исследованиях. М., 1989.
4. *Сатаров Г. А.* Многомерное шкалирование // Интерпретация и анализ данных в социологических исследованиях. М., 1989.
5. *Гласс Дж., Стенли Дж.* Статистические методы в педагогике и психологии : пер. с англ. М. : Прогресс, 1976.
6. *Мандель И. Д.* Кластерный анализ. М. : Финансы и статистика, 1988.
7. *Poole K. T., Rosental H. A.* A Spatial Model for Legislative Rallcall Analysis // American Journal of Political Science. 1985. V. 29, № 2. P. 357–384.
8. *Sattath S., Tversky A.* Additive Similarity Trees // Psychometrika. 1977. V. 42, № 7. P. 319–345.

СОЦИАЛЬНЫЕ ДЕТЕРМИНАНТЫ САМООЦЕНКИ УСПЕХА[1]

С. В. Сивуха, М. Титма

Исследования качества жизни и субъективного благополучия в поведенческих науках всегда отличались практической значимостью, поскольку результаты непосредственно влияли на социальную политику. Эта взаимосвязь обеспечивала финансирование конкретных исследова-

[1] Опубликовано: Социальное расслоение возрастной когорты. Выпускники 80-х в постсоветском пространстве. С. 250–274.

ний и в итоге способствовала развитию концептуальных основ направления. В последние десятилетия теоретические подходы к исследованию субъективного благополучия стали глубже, отражая реальную сложность изучаемых процессов. Одной из основных теоретических проблем сегодня является довольно слабое соответствие между объективными и субъективными мерами качества жизни или благополучия человека. Многие исследователи продемонстрировали в своих работах, что поведение индивидов и групп более успешно объясняется и предсказывается именно с помощью субъективных индикаторов, хотя и объективные меры благополучия по-прежнему задействуются (преимущественно в межстрановых сравнительных исследованиях). Показательно также, что не только психологи, но и экономисты практикуют определение благополучия в субъективном ракурсе [13, 14]. Поэтому использование субъективных показателей при исследовании жизненного пути представляется вполне закономерным.

В настоящей главе мы рассматриваем самооценку молодыми людьми своих возможностей (или, иначе, реального успеха на данном жизненном этапе) в условиях резкого изменения общества. Для этого мы используем данные шести регионов бывшего Советского Союза. Вначале излагаются исследовательский подход, методические вопросы, затем представляются результаты и дискуссия по выдвинутым гипотезам.

В отличие от других исследований, выполненных на российском материале [1, 2], наша работа основывается на тщательно эксплицированном теоретическом подходе, строится на сравнительном анализе, включает строгую проверку выдвинутых гипотез.

Исследовательский подход

Субъективная оценка жизни как предмет психологического и социологического исследования имеет давние традиции. В качестве объекта исследования используются качество жизни, удовлетворенность жизнью в целом и ее различными областями, субъективное благополучие, счастье. Многие авторы сходятся на идее о концептуальной близости этих понятий, в пользу чего говорит постоянно отмечаемая статистическая связь между соответствующими эмпирическими показателями [5, 6, 12].

Выбор конкретных субъективных мер благополучия определяется не только теоретической ориентацией, но в значительной степени (в том числе на эмпирическом уровне) остается делом личного вкуса исследователя. Пространство оценок задается следующими осями: от интегральных, целостных оценок (удовлетворенность жизнью) до конкретных показателей (удовлетворенность браком); от когнитивных измерений (удовлетворенность жизнью) до аффективных (счастье, самочувствие); от внешних референтов до внутренних референтов. Внешние референты предполагают соотнесение результатов индивида с результатами нормативной выборки

и чаще используются в методиках тестового характера, внутренние референты требуют соотнесения с индивидуальными стандартами. В свою очередь, использование внутренних референтов может носить характер социального (с другими) либо индивидуального сравнения (с прошлым, желаемым или идеальным состоянием индивида).

В исследованиях с использованием субъективных мер благополучия, таких как удовлетворенность жизнью и счастье, наиболее влиятельными являются две теории (развернутое обсуждение предположений, лежащих в основе этих теорий, см. в [22]). В теории сравнения предполагается, что основой для суждений индивида о своем благополучии является относительная депривация потребностей; индивидуальная оценка жизни взвешивается в соответствии с некоторыми стандартами. Базовыми для субъективной оценки благополучия процессами являются социальное сравнение [11] или сравнение желаемого и достигнутого.

Теория условий жизни (livability) и родственные ей теории базовых потребностей и ресурсов основываются на предположении, что субъективное благополучие зависит от объективного качества жизни, от доступности ресурсов и удовлетворения основных потребностей.

В последние годы теория сравнения получила большую популярность не только в силу эмпирических свидетельств, но и в связи с развитием понятийной рамки, в особенности в теории множественных сравнений (multiple-discrepancies theory [18]). Эта когнитивная теория, базирующаяся на нескольких психологических концепциях, утверждает, что удовлетворенность жизнью в целом и ее отдельными сторонами объясняется множественными сравнениями между тем, что человек имеет, и тем, что он хотел бы иметь, что имеют другие, чего индивид ожидает в будущем или связывает с лучшим возможным положением и т. д., некоторые исследователи пришли к выводу, что именно модель множественных сравнений лучше всего соответствует эмпирическим данным [9, 19].

Объяснительные возможности теории множественных сравнений еще больше повышаются, когда в структурных моделях рассматриваются двусторонние связи (восходящие и нисходящие) между удовлетворенностью различными сторонами жизни [16, 19]. Это обстоятельство используется для конструктивной валидизации теории, хотя, на наш взгляд, оно не может быть серьезным аргументом в пользу теории множественных сравнений. Связи между интегральными и частными индикаторами благополучия могут быть объяснены феноменом когнитивной согласованности [3], поскольку удовлетворенность жизнью по существу является когнитивной оценкой [17]. Меры удовлетворенности различными сторонами жизни являются не причинами и следствиями общей удовлетворенности, но, скорее, ее составляющими.

Аргументом в пользу теории множественных сравнений не может служить и высокая корреляция между общей удовлетворенностью и несоот-

ветствием между «имею/хочу иметь». Как отмечает В. Шульц, речь здесь может идти об искусственной связи, порождаемой близостью латентных переменных. Следует также отметить, что доля дисперсии, объясняемой моделями сравнения, значительно увеличивается, если учитываются явно психологические переменные (самооценка). В этом случае речь может идти о преимущественном внимании к изучению феноменов сознания, но не социально-экономических детерминант благополучия человека. Таким образом, теория сравнения не свободна от концептуальных проблем.

Теория условий жизни (базовых потребностей), в значительной степени опирающаяся на здравый смысл, утверждает, что удовлетворенность жизнью зависит не от относительного, а от абсолютного удовлетворения потребностей [8, 22]. Люди, живущие в лучших условиях, должны быть в большей степени удовлетворены жизнью. Оценка этих условий как пригодных для жизни зависит от возможности общества удовлетворить индивидуальные потребности. Одно из важнейших предположений – это идея основных потребностей. Существенно, что никаких других предположений не делается. Теория предсказывает, что различия в дисперсии удовлетворенности жизнью в разных странах связаны с социальным неравенством. Не случайно, что в исследованиях в этом русле используются межстрановые сравнения [8, 22].

Мы рассматриваем теорию условий жизни как весьма перспективную для исследования нашей социальной реальности. Во-первых, объяснительные возможности теории достаточно высоки. Во-вторых, ее отличает относительная простота, поскольку не требуется специальных предположений о процессах субъективной оценки жизни. В-третьих, теория особенно перспективна в исследуемых обществах, где нарастают социальное расслоение и поляризация условий жизни, а темпы этого расслоения в различных странах бывшего Советского Союза различны. В-четвертых, установление внеличных объективных причин (условий) важно для управления социальной политикой в новых независимых государствах.

В защиту теории условий жизни говорят также исследования влияния дохода на удовлетворенность жизнью. Это влияние поддается абсолютному, а не относительному объяснению [8, 21]. Хотя эта связь нелинейна, она, как было показано, все же сохраняется в социальных группах с высокими доходами.

В. Шульц [19] полагает, что теория ресурсов – один из вариантов теории базовых потребностей – может быть усовершенствована за счет эксплицитного включения когнитивных процессов. С другой стороны, она требует представления объективных условий в терминах ресурсов таким образом, чтобы демографические характеристики (пол) были заменены стоящими за ними ресурсными переменными (образование, доходы).

Мы не разделяем предложения В. Шульца об исключении из анализа демографических переменных. Несмотря на то, что в ряде исследова-

ний не было установлено значимых связей между демографическими показателями и удовлетворенностью жизнью, мы намерены рассмотреть эти переменные в своих моделях прежде всего потому, что в условиях усиливающегося неравенства и трансформаций они должны иметь некоторое значение. Кроме того, взаимодействие между собственно ресурсными переменными, такими как доход, образование, и демографическими (пол) в наших обществах изучено недостаточно.

В настоящем исследовании мы ставили целью идентифицировать объективные условия жизни и ресурсы (независимые переменные), влияющие на субъективное благополучие (зависимую переменную). Задача структурного моделирования взаимоотношений между общей оценкой качества жизни и удовлетворенностью отдельными сторонами жизни остается за рамками данной публикации.

Чтобы избежать множественности зависимых переменных и при этом сохранить многомерность субъективной оценки жизненных возможностей, мы выделяем типы респондентов в соответствии с их удовлетворенностью различными аспектами жизни. Информация об индивидуальных возможностях удовлетворения потребностей в различных областях жизни по сравнению со сверстниками используется для конструирования новой номинальной зависимой переменной, фиксирующей различные констелляции индивидуальных возможностей. В этой части наше исследование носит эксплораторный характер.

Затем мы конструируем мультиноминальную модель, подобную регрессионной, и оцениваем влияние некоторых условий жизни на тип субъективной оценки индивидуальных возможностей. Теоретической основой моделей является синтез рассмотренных ранее идей множественного сравнения, социального сравнения, ресурсов. В этой части мы проверим на своей модели несколько содержательных гипотез.

Мы исходим из того, что в период трансформации доступность ресурсов, объективные условия жизни, сами потребности и их относительная значимость, а также стандарты восприятия и когнитивная оценка условий жизни у разных социально-экономических групп в разных областях жизни изменились в разной степени. В этой связи мы сравниваем мультиноминальные модели для двух групп регионов бывшего Советского Союза, различающихся по темпам рыночных реформ (Балтийские страны, с одной стороны, и Свердловская, Карагандинская область и Краснодарский край – с другой).

Гипотезы исследования

Развитие рыночных отношений приводит к тому, что социально-экономические факторы в большей степени поляризуют возможности удовлетворения молодежью своих потребностей. Речь идет не о том, что близость общества к рынку автоматически расширяет пространство инди-

видуальных возможностей, а скорее о том, что субъективное благополучие человека становится в большей степени зависимым от социально-экономических факторов. Поэтому мы ожидаем, что независимые переменные объяснят большую долю дисперсии зависимой переменной в модели для Балтийских стран.

Гипотеза 1. Влияние социально-экономических переменных на оценку респондентами своих возможностей более заметно в странах с более развитыми рыночными отношениями.

Несмотря на то, что западные исследователи отмечают отсутствие связи между полом и субъективными мерами качества жизни [10, 15, 19], мы полагаем, что в условиях социальных и экономических преобразований мужчины имеют некоторые преимущества в силу их большей мобильности и лучших возможностей в выборе работы.

Гипотеза 2. Мужчины имеют бóльшие, по сравнению с женщинами, возможности, они с большей вероятностью попадут в кластеры «успешных» и с меньшей вероятностью — в кластеры «неуспешных».

Факторы жизненного цикла двойственно влияют на возможности индивидов. Наличие несовершеннолетних детей «связывает» индивидов и уменьшает их возможности. Семья также ограничивает возможности индивидов, в особенности жить независимо, по своему усмотрению. Проживание в областном городе или столице может иметь разнонаправленные эффекты. Диапазон потенциальных возможностей в шансах иметь интересную работу и жить по своему усмотрению, в крупных, по сравнению со средними, городах шире, однако и социальное расслоение в столицах более ощутимо. Кроме того, существенно, что среда большого города приводит к завышению стандартов сравнения и занижению оценок респондентов своих возможностей. Таким образом, у жителей региональных центров более вероятна негативная оценка своих возможностей. Проживание в сельской местности ограничивает возможности населения. Эти соображения приводят к следующим двум гипотезам.

Гипотеза 3. Наличие детей уменьшает возможности индивидуального успеха. Наличие семьи будет иметь отрицательный эффект на вероятность принадлежности к группе «независимых».

Гипотеза 4. Проживание в столице или областном центре повышает вероятность попадания в группы «успешных» и «независимых». Эффект проживания в региональных центрах на попадание в группы «успешных» ожидается нулевой (незначимый). Проживание в сельской местности связано с низкой оценкой респондентами своих возможностей.

Образование, являясь важным социальным ресурсом, должно положительно влиять на субъективные оценки качества жизни. В соответствии с теорией человеческого капитала, в обществах с более развитыми рыночными отношениями влияние уровня образования на качество жизни является более сильным.

Г и п о т е з а 5. Среднее специальное и высшее образование в сравнении с общим средним образованием расширяет возможности индивидов. Эта закономерность будет более заметной в Балтийских странах.

Положение на рынке труда, как мы ожидаем, также сильно влияет на возможности индивидов. Самозанятость (предпринимательство и индивидуальная трудовая деятельность) по сравнению с наемной работой в течение полного дня будет расширять возможности респондентов, а статус безработного будет иметь отрицательный эффект. В отношении занятости в домашнем хозяйстве трудно ожидать однозначного эффекта с содержательной точки зрения, поскольку в эту категорию входят и те, кто имеет высокий достаток и, следовательно, хорошие возможности для удовлетворения материальных и духовных потребностей, и те, кто занят в домашнем хозяйстве вынужденно. В отношении индивидов, находящихся вне рынка труда, у нас нет содержательных предположений. Оценивая свои возможности, эти респонденты могут принимать во внимание свой временный статус и соответствующим образом выбирать группы сравнения или делать оптимистические оценки компенсаторного характера.

Г и п о т е з а 6. Респонденты, занятые предпринимательством, будут оценивать свои возможности высоко, а безработные – низко. Занятость в домашнем хозяйстве будет иметь двойной эффект: вероятность отнесения этих индивидов и к группе «успешных», и к группе «неуспешных» будет выше по сравнению с вероятностью попадания в референтную группу. Положение вне рынка труда, как ожидается, не будет иметь эффекта на оценки респондентами своих возможностей.

Влияние доходов и материального достатка на возможности индивидов, на наш взгляд, достаточно предсказуемо. Разумно предположить, что первый квинтиль доходов существенно ограничивает эти возможности, а пятый квинтиль – расширяет. Обладание легковым автомобилем должно иметь тот же эффект.

Г и п о т е з а 7. Первый квинтиль доходов уменьшает вероятность принадлежности респондентов к группе с хорошими возможностями («успешных») и увеличивает вероятность отнесения к группе с ограниченными возможностями («неуспешных»). Пятый квинтиль доходов и обладание легковым автомобилем имеют противоположный эффект.

Социальные страты не будут оказывать такого сильного дифференцирующего влияния на возможности респондентов. Значимые эффекты будут иметь страты, в которых образовательный уровень и квалификация позволят быстро адаптироваться к изменившейся социальной реальности через трудовую мобильность, в частности, через переход в частный сектор экономики или в государственный аппарат, через работу по совместительству и т. д. К таким стратам следует отнести руководителей и профессионалов. Важен также другой фактор – близость к сферам распределения

общественных благ. Система распределения, остаточные формы которой унаследованы от социалистической системы хозяйствования, продолжала играть довольно важную роль в 1992—1993 гг. В этой связи можно ожидать, что руководители, конторские служащие (в том числе государственные и банковские) и работники сферы обслуживания получат по сравнению с референтной стратой (рабочими) некоторые преимущества. Аграрные рабочие объективно имеют меньше возможностей, чем индустриальные, однако в соответствии с теорией сравнения они сопоставляют свое положение с представителями своей собственной, относительно гомогенной группы, и, таким образом, малоправдоподобна низкая оценка ими своих возможностей. Из этих соображений следует наша следующая гипотеза.

Гипотеза 8. Руководители, профессионалы, конторские служащие и работники сферы обслуживания будут больше представлены в группах с широкими возможностями, меньше — в группах с ограниченными возможностями.

Работа в частном секторе по сравнению с работой в секторе, финансируемом государством, как можно ожидать, предоставляет индивидам широкие возможности и одновременно защищает их от депривации своих потребностей. Это утверждение можно отнести к части работников госбюджетного сектора, а именно к служащим государственного аппарата. Однако значительная часть госбюджетных работников (наука, культура, здравоохранение) пользуется лишь вторым из этих преимуществ, и значимые эффекты следует ожидать лишь на одном из полюсов пространства возможностей.

Гипотеза 9. Вероятность попадания в группу «успешных» будет выше для работающих в частном секторе. Вероятность отнесения к группе «неуспешных» будет ниже для работников частного и госбюджетного секторов.

Методические вопросы измерения

Респондентов просили оценить по 4-балльной шкале, насколько удачно складывалась их жизнь в целом. Затем им задавали 6 вопросов об удовлетворенности жизнью: «Исходя из чего Вы назвали Вашу жизнь удачной (неудачной)? Сопоставьте Вашу жизнь с жизнью Ваших сверстников. Каковы Ваши возможности по сравнению со сверстниками: иметь интересную работу; приобретать ценные вещи; жить по своему усмотрению; хорошо питаться; покупать модную одежду; продвигаться по службе?». Эти оценки давались по 5-балльной шкале, от 1 (мои возможности значительно меньше) до 5 (мои возможности значительно больше). Все вопросы задавались в конце биографического интервью.

Такой дизайн интервью, несомненно, должен влиять на характер ответов. Полагаем, что размещение вопросов об удовлетворенности жизнью

в конце интервью (после биографической рефлексии) способствовало увеличению искренности и взвешенности ответов и уменьшению влияния факторов социальной желательности и социальных норм [23]. Оценки респондентами своих возможностей по сравнению со сверстниками, строго говоря, не являются субъективными оценками благополучия; их следует квалифицировать как оценки доступности социальных ресурсов в разных областях жизни, полученные посредством эксплицитного социального сравнения (domain-specificself/others discrepancies, [18]). С точки зрения теории множественных сравнений эти оценки являются хорошими предсказателями субъективного благополучия. Кроме того, удовлетворяется требование В. Шульца [19] о введении когнитивных операций (сравнений) в теорию ресурсов. Таким образом, речь идет о конструктивном синтезе обеих рассмотренных теорий.

Мы использовали эти оценки возможностей для конструирования номинальной переменной «Жизненный успех». С этой целью мы провели иерархический кластерный анализ респондентов по методу Уорда отдельно для каждой группы респондентов. В качестве меры близости использовались квадраты евклидовых расстояний. Поскольку оценки респондентами своих возможностей хорошо питаться оказались недискриминативными, а мера возможностей служебного продвижения имела большое количество отсутствующих значений (неработающие респонденты часто не давали ответа на этот вопрос), в кластерном анализе использовались четыре переменные: возможности иметь интересную работу, приобретать ценные вещи, жить по своему усмотрению и покупать модную одежду. Было рассмотрено несколько кластерных решений. Мы остановились на шестикластерном, поскольку оно оказалось дискриминативным и хорошо интерпретируемым, причем интерпретация кластеров в обоих массивах данных была довольно согласованной, и доли респондентов, отнесенных к соответствующим кластерам, оказались пропорциональными. Принадлежность к кластеру в шестикластерном решении в дальнейшем анализе использовалась в качестве зависимой переменной. В табл. 1 приведены характеристики кластеров для каждого из массивов. Нумерация кластеров упорядочена по степени убывания удовлетворенности жизнью в целом.

Интерпретация кластеров в обеих группах респондентов довольно близка. Первый кластер – люди с хорошими возможностями во всех областях («успешные»). Второй – респонденты с довольно хорошими возможностями, в особенности в том, что касается работы («довольно успешные»). Третий, самый многочисленный кластер объединил людей с умеренными возможностями («средние»). В четвертый вошли респонденты, которые оценили возможности иметь интересную работу и жить независимо довольно высоко, а материальные возможности – низко («независимые»). Пятый кластер объединил респондентов с малыми

возможностями, особенно в том, что касается возможности иметь интересную работу («неуспешные в работе»). В шестой кластер вошли индивиды с умеренными возможностями иметь интересную работу и низкими возможностями во всех остальных сферах («неуспешные вне работы»). Поскольку различение двух первых и двух последних кластеров в нашем исследовании с содержательной точки зрения не представляет интереса, далее в тексте они выступают под общими именами – соответственно «успешные» и «неуспешные».

Таблица 1

Средние значения возможностей респондентов, самооценок успешности жизни по кластерам, наполненность кластеров

Переменная	Регионы	Кластер					
		1	2	3	4	5	6
		успешные		референт-ные	независи-мые	неуспешные	
Интересная работа	Балтия	3,42	3,79	3,05	3,31	1,77	2,51
	СНГ	4,24	2,94	3,05	3,24	1,85	2,42
Ценные вещи	Балтия	3,84	3,15	2,86	1,79	2,26	1,74
	СНГ	3,47	3,40	3,03	1,83	2,44	1,69
Самостоятель-ность	Балтия	3,62	3,69	2,85	3,46	3,34	1,71
	СНГ	3,63	3,87	2,83	3,30	2,89	1,71
Модная одежда	Балтия	3,84	3,01	2,90	1,94	2,54	1,70
	СНГ	3,48	3,54	2,81	2,17	2,58	1,88
Успех в жизни	Балтия	3,01	2,98	2,92	2,84	2,77	2,68
	СНГ	2,97	2,90	2,92	2,78	2,66	2,52
Количество случаев	Балтия	655	723	1300	626	232	733
	СНГ	698	537	1191	726	279	503

В исследовании использованы данные 6 регионов, для целей анализа объединенных в две относительно гомогенные и различимые группы. В первую вошли Балтийские страны (4844 респондента), во вторую – два региона России (Свердловская область и Краснодарский край) и Карагандинская область Казахстана (4364 случая). Регионы объединены по критериям уровня жизни, темпов социально-экономических преобразований, культурных особенностей. После удаления отсутствующих значений по всем переменным в мультиноминальной модели для Балтийских стран оказалось 3838 случаев, в модели для второй группы стран – 3316 случаев.

В качестве независимых переменных мы использовали три набора факторов: ориентированные на достижения (образование, доход, положе-

ние на рынке труда, сектор экономики, социальная страта), аскриптивные (пол) и факторы жизненного цикла (семейный статус, наличие детей, тип населенного пункта). Это различение сделано по принципу управляемости. Первая группа факторов в большей степени подвластна контролю со стороны респондентов, аскриптивные факторы индивидуально неуправляемы, последняя группа факторов связана с жизненными циклами и в значительной степени подвержена влиянию индивидуальной истории и культурных норм. Тип населенного пункта мы отнесли к факторам жизненного цикла, поскольку место жительства респондентов в советских и постсоветских условиях определялось не только возможностями получения прописки и местом жительства родителей, но также образовательной миграцией и вступлением в брак.

Все переменные были преобразованы в дихотомические. Например, пол кодирован «1» для мужчин и «0» для женщин, поэтому дихотомическая переменная получила название «Мужчины». В переменной «Семья» статусы холостых/незамужних, разведенных и вдовых кодированы нулем.

Тип населенного пункта представлен тремя дихотомическими переменными: столица, областной центр; большие и средние города; малые города, поселки, села. Положение респондентов на рынке труда представлено следующим набором дихотомических переменных: занятые полный день; занятые неполный день и работающие по срочному договору; самозанятые; находящиеся вне рынка труда (по состоянию здоровья, в связи с декретным отпуском, учащиеся, проходящие переобучение); безработные; домохозяева.

Социальная страта респондентов определялась по модифицированной шкале профессионального статуса ISCO. В силу особенностей кодировки данных после предварительного их изучения мы решили не проводить различения между квалифицированными и неквалифицированными рабочими. Таким образом, в исследовании представлены следующие социальные страты: руководители, профессионалы (с высшим образованием), полупрофессионалы (со средним специальным образованием), конторские служащие, работники сферы обслуживания и торговли, промышленные рабочие, сельскохозяйственные рабочие. Сектор экономики кодировался так: госбюджетный, финансируемый государством (госхозрасчетный), частный, другое.

Доходы респондентов были преобразованы в квинтили. В условиях различий денежных единиц в разных странах это позволило получить сопоставимые оценки. Переменная «Автомобиль» фиксирует наличие легкового автомобиля в собственности респондента и/или его семьи. В табл. 2 представлены описательные статистики независимых переменных. Знаком «*» отмечены референтные категории в наборах дихотомических переменных, представляющих номинальные переменные.

Таблица 2

Описательные статистики дихотомизированных независимых переменных

Переменная	Среднее	Переменная	Среднее
ПОЛ		ДОХОД	
Мужчины	0,47	Квинтиль 1	0,19
СЕМЬЯ		Квинтиль 2	0,21
есть	0,75	Квинтиль 3*	0,20
ДЕТИ		Квинтиль 4	0,19
есть	0,75	Квинтиль 5	0,20
ТИП ПОСЕЛЕНИЯ		Автомобиль	0,26
Обл. центр	0,30	СОЦИАЛЬНАЯ СТРАТА	
Город*	0,38	Руководители	0,06
Село, поселок	0,32	Профессионалы	0,17
ОБРАЗОВАНИЕ		Полупрофессионалы	0,19
ПТУ	0,24	Конторские служащие	0,08
Общее среднее*	0,17	Сервис	0,09
Среднее спец.	0,37	Пром. рабочие *	0,38
Высшее	0,22	Сельхоз. рабочие	0,03
ЗАНЯТОСТЬ		СЕКТОР ЭКОНОМИКИ	
Полнозанятые*	0,66	Госбюджетный	0,43
Временно занятые	0,05	Госфинансируемый	0,34
Самозанятые	0,03	Частный	0,13
Неработающие	0,15	Другое	0,09
Безработные	0,06		
Домохозяева	0,05		

Для оценки каузальных моделей, где в качестве независимых переменных выступали социальные и экономические ресурсы и условия жизни, а в качестве зависимой переменной – кластеры социальных и экономических возможностей респондентов (номинальная переменная), мы использовали мультиноминальный логистический регрессионный анализ [4], в котором логарифм отношения вероятности попадания в некоторую группу к вероятности попадания в референтную группу представляется как линейная функция вектора независимых переменных. Регрессион-

ный коэффициент при независимой переменной для некоторой *j*-й категории зависимой переменной интерпретируется следующим образом: отношение вероятности попадания в *j*-ю группу к вероятности попадания в референтную группу при прочих равных условиях равно числу *e*, возведенному в степень, равную регрессионному коэффициенту. В качестве референтной категории зависимой переменной по содержательным и формальным соображениям был выбран кластер умеренных возможностей. Оценка моделей проведена с помощью программы RATE [19].

Для того чтобы оценить относительное влияние отдельных независимых переменных и групп переменных на зависимую переменную, мы использовали подход вложенных моделей. Первая вложенная модель включала аскриптивный фактор (пол) и факторы жизненного цикла, во вторую модель были добавлены образование и положение на рынке труда, в третью – показатели материального благосостояния, в четвертую – социальная страта и сектор экономики.

Результаты и обсуждение

Продвинутость рыночных реформ в различных регионах бывшей Советской страны самым непосредственным образом сказывается на субъективной оценке качества жизни населения. Удовлетворенность жизнью среди молодежи Балтийских стран значимо выше, чем в регионах России и Казахстана ($t = 5{,}38$, $p < 0{,}001$).

Рассмотрим совокупную оценку значимости независимых переменных в мультиноминальных моделях для двух массивов данных (табл. 3).

Таблица 3

Псевдо-*R*-квадрат вложенных регрессионных моделей

Группа независимых переменных	Балтийские страны	Русскоязычные регионы СНГ
Пол, факторы жизненного цикла	0,174	0,158
Образование, занятость	0,193	0,167
Доход, автомобиль	0,196	0,127
Страта, сектор экономики	0,087	0,108
Всего	0,634	0,549

Псевдо-*R*-квадрат интерпретируется как показатель улучшения мультиноминальной модели после включения независимых переменных. В целом социально-экономические факторы оказались более значимыми в предсказании субъективной оценки респондентами своих возможностей в Балтийских странах. Эта закономерность особенно явственна в отношении переменных материального благосостояния. Данные субъектив-

ной оценки респондентами своих возможностей лучше объясняются социально-экономическими переменными на балтийском массиве. Первую гипотезу можно считать подтвержденной.

Обратимся к проверке других содержательных гипотез. В табл. 4 приведены регрессионные коэффициенты и показатели качества оцененных моделей со всеми независимыми переменными. Группировка переменных соответствует порядку включения переменных в модель. Все коэффициенты Хи-квадрата (разность между показателями максимального правдоподобия последовательных вложенных моделей, умноженная на 2) значимы на уровне $p < 0{,}0001$, что является эмпирическим подтверждением истинности построенных моделей.

Мужчины, по сравнению с женщинами, представлены в кластерах «неуспешных» в меньшей степени, чем в референтном кластере («умеренные успешные») в обеих группах регионов. Другими словами, в исследованных регионах мужчины в меньшей степени сталкиваются с ограничениями индивидуальных возможностей. Однако представленность мужчин в кластерах «успешных» оказалась такой же, как и в референтной группе, или даже ниже (в российско-казахстанском массиве). Значительная доля респондентов мужского пола вошла в референтный кластер – группу индивидов с умеренными социальными возможностями. Возможно, это объясняется тем, что мужчины дают более осторожные оценки своим возможностям и в меньшей степени подвержены влиянию фактора защитного оптимизма. Поскольку ни одна психологическая переменная в оцененных моделях не контролировалась, мы не имеем возможности проверить это утверждение. Таким образом, в отношении второй гипотезы получены смешанные свидетельства.

Гипотеза о том, что наличие несовершеннолетних детей уменьшает возможности индивидов, подтвердилась в обеих группах регионов. Респонденты, имеющие детей, с меньшей вероятностью попадали в кластеры «успешных» и с большей вероятностью – в группу «неуспешных». Индивиды, состоящие в браке, реже представлены в кластерах «независимых». Более того, в обеих моделях все эффекты семейного статуса отрицательны. Таким образом, индивиды, состоящие в браке, реже дают крайние оценки своим возможностям. В российско-казахстанском массиве значимы четыре регрессионных коэффициента из пяти: семья, при прочих равных условиях, ограничивает жизненные возможности индивидов, но в то же время выполняет компенсаторную функцию, защищая от переживания неблагополучия. Возможно, это связано с традиционными ценностями жителей российской провинции. С другой стороны, отсутствие подобной связи в балтийском массиве можно интерпретировать как следствие маркетизации общества, в котором вес семьи в оценке субъективного благополучия уменьшается.

Таблица 4

Регрессионные коэффициенты для мультиномиальной модели

Независимая переменная	Регионы	Кластер				
		1	2	4	5	6
		«успешные»		независимые	«неуспешные»	
Константа	Балтия	-0,96**	-0,96**	-0,27	-1,19**	-0,31
	СНГ	-0,91**	0,14	-0,11	-0,20	-0,80**
Пол: мужчины	Балтия	-0,24	-0,05	-0,18	-0,74**	-0,47**
	СНГ	0,07	-0,40**	-0,23	-0,60**	-0,51**
Семья	Балтия	-0,04	-0,16	-0,37*	-0,28	-0,25
	СНГ	-0,43*	-0,23	-0,53**	-0,52*	-0,71**
Дети	Балтия	-0,46**	-0,43**	-0,02	0,37	0,44**
	СНГ	-0,14	-0,47**	0,01	-0,00	0,70**
ТИП ПОСЕЛЕНИЯ						
Областной центр	Балтия	-0,12	0,17	0,28*	0,02	0,28*
	СНГ	-0,17	-0,55**	-0,31*	-0,33	0,15
Поселок	Балтия	-0,49**	-0,29*	0,05	0,36*	-0,08
	СНГ	0,12	-0,54**	-0,24	0,02	0,24
ОБРАЗОВАНИЕ						
ПТУ	Балтия	0,28	0,37	0,36	0,15	0,25
	СНГ	-0,10	-0,30	0,22	-0,23	-0,06
Среднее специальное	Балтия	0,06	0,51**	0,17	-0,27	0,12
	СНГ	0,23	-0,12	0,04	-0,69**	-0,03
Высшее	Балтия	0,49*	0,81**	0,44	-0,52	-0,25
	СНГ	0,53*	-0,09	0,13	-0,87*	-0,62*
ЗАНЯТОСТЬ						
Временно занятые	Балтия	0,22	-0,24	-0,20	0,19	-0,15
	СНГ	0,86**	0,69	0,62	0,65	0,55
Самозанятые	Балтия	0,88**	0,43	-0,62	-7,97	-0,66
	СНГ	0,22	1,01**	-0,17	0,45	-0,01
Вне рынка	Балтия	0,27	-0,18	-0,27	0,37	0,12
	СНГ	-0,11	-0,05	0,01	0,14	0,16

Продолжение табл. 4

Независимая переменная	Регионы	Кластер				
		1	2	4	5	6
		«успешные»		независимые	«неуспешные»	
Безработные	Балтия	0,00	-0,63*	-0,36	0,77**	0,31
	СНГ	-0,04	0,67*	0,43	0,76*	0,47
Домохозяева	Балтия	0,79**	-0,19	-0,43	0,52	0,17
	СНГ	-0,27	0,23	-0,47	-0,17	-0,38
ДОХОД						
Квинтиль 1	Балтия	-0,01	-0,08	0,02	-0,14	-0,10
	СНГ	0,17	-0,23	0,23	0,18	0,40*
Квинтиль 2	Балтия	0,16	0,06	0,05	-0,23	0,04
	СНГ	0,12	-0,29	0,21	-0,12	0,55**
Квинтиль 4	Балтия	0,27	0,19	-0,19	-0,62*	-0,29
	СНГ	0,25	0,11	0,15	-0,27	0,23
Квинтиль 5	Балтия	0,75**	0,64**	-0,71**	-0,29	-0,48**
	СНГ	0,51**	0,11	-0,15	-0,42	-0,46*
Автомобиль	Балтия	0,46**	-0,06	-0,53**	-0,20	-0,58**
	СНГ	0,56**	0,59**	-0,27	-0,22	-0,26
СОЦИАЛЬНЫЙ СТАТУС						
Руководители	Балтия	0,69**	0,58**	-0,15	-0,55	-0,35
	СНГ	1,18**	0,20	0,20	0,35	-0,15
Профессионалы	Балтия	-0,27	0,42	0,03	-0,07	-0,10
	СНГ	0,27	-0,46	0,25	-0,46	0,02
Полупрофессио-налы	Балтия	0,15	0,23	0,11	0,08	-0,19
	СНГ	0,65**	0,31	0,36*	-0,24	-0,00
Служащие	Балтия	0,50*	0,50*	0,12	0,15	0,16
	СНГ	0,13	-0,04	-0,14	0,20	-0,21
Обслуживание	Балтия	0,30	-0,16	-0,10	-0,04	-0,51*
	СНГ	0,19	0,35	0,10	0,11	-0,26
Аграрные рабочие	Балтия	-0,34	0,25	-0,03	0,23	0,33
	СНГ	0,49	0,70	0,23	0,97	0,43
СЕКТОР ЭКОНОМИКИ						
Госбюджетный	Балтия	-0,08	-0,03	0,04	-0,39*	0,24
	СНГ	-0,11	-0,03	0,12	-0,07	0,02
Частный	Балтия	0,02	0,12	-0,27	-0,61*	-0,34
	СНГ	0,41*	0,53**	-0,16	-0,18	-0,02

Окончание табл. 4

Независимая переменная	Регионы	Кластер				
		1	2	4	5	6
		«успешные»		независимые	«неуспешные»	
Другой	Балтия	0,15	0,05	-0,03	0,04	-0,18
	СНГ	0,40	0,40	0,61*	-0,53	0,40
Максимальное правдоподобие	Балтия					-6080
	СНГ					-5339
Количество случаев	Балтия					3838
	СНГ					3316

Примечание. Уровни значимости регрессионных коэффициентов обозначены: $*p < 0{,}05$; $**\,p < 0{,}01$.

Проживание в столицах имеет предсказанные эффекты в Балтийских странах. Жители региональных центров российской и казахстанской провинции оценивают свои возможности невысоко, их доля в кластерах «успешных» и «независимых» значимо меньше, чем в референтной группе. В отличие от жителей балтийских столиц, жители областных центров меньше представлены в группе «неуспешных», что косвенно подтверждает наше предположение о влиянии степени социального расслоения на субъективную оценку благополучия. Жители сельской местности воспринимают свои возможности как более ограниченные, и эта тенденция сильнее проявляется в Балтийских странах (три коэффициента значимы, и их знак совпадает с предсказанным).

Данные табл. 4 дают важные свидетельства в пользу предположения о положительном влиянии среднего специального и высшего образования (по сравнению с общим средним) на субъективную оценку респондентами своих возможностей. Любопытно, что сила этого влияния на разных полюсах пространства возможностей в двух исследованных массивах респондентов оказалась разной. В Балтийских странах образование повышает вероятность попадания в «успешные» кластеры (позитивный, деятельный эффект). В российско-казахстанском массиве среднее специальное и высшее образование снижает вероятность попадания в кластеры «неуспешных» (защитное, пассивное влияние). Первый эффект прямо интерпретируется в терминах теории человеческого капитала, тогда как «защитная» роль образования в российско-казахстанской модели является, на наш взгляд, рудиментом дорыночного общественного устройства. Мы усматриваем в этих результатах еще одно важное доказательство того, что связь между социально-экономическими факторами и субъективным благополучием опосредована степенью развитости рыночных отношений.

Самозанятость, как мы и предполагали, в целом расширяет возможности респондентов. Знаки регрессионных коэффициентов в балтийском массиве совпали с предсказанными, значимым оказался единственный коэффициент. В Свердловске, Краснодаре и Караганде коэффициент одного из кластеров «неуспешных» имеет знак, обратный предсказанному. Хотя эффект этот незначим, полезно вспомнить в этой связи дискуссию в российской социологической литературе по поводу сложности группы самозанятых, которая объединяет успешных предпринимателей, «челноков» и людей, берущихся за любое дело, чтобы свести концы с концами.

Положение вне рынка труда не оказало влияния на оценки респондентами своих возможностей. Безработные в балтийском массиве значимо чаще попадали в кластер «неуспешных» (пятый) и значимо реже – в кластер «успешных» (второй). Однако регрессионные коэффициенты для полярных кластеров оказались незначимо отличными от нуля. Более того, в российско-казахстанском массиве безработные существенно чаще были представлены в кластере «успешных», что противоречит выдвинутой гипотезе. Возможно, это связано с номинальным характером сконструированной зависимой переменной (артефакт измерения). Но поскольку другие утверждения шестой гипотезы подтверждены, мы склоняемся ко второму возможному объяснению этих эффектов. Безработица, как и самозанятость, является более сложным явлением, чем это представляется обыденному сознанию. В частности, статус безработного может быть прикрытием для самостоятельной экономической активности индивидов. Существенно, что подобный результат отмечен только в обществах со слаборазвитыми рыночными отношениями.

Доходы и такой показатель материального достатка, как автомобиль в личной собственности, оказались наилучшими предикторами субъективного благополучия. Знаки регрессионных коэффициентов и их общий рисунок соответствуют предсказанным для обоих массивов данных. Более дифференцирующими оказались пятый квинтиль доходов и наличие автомобиля. Вероятно, это связано с тем, что респонденты с высоким достатком ориентируются на весьма отличные референтные группы. Следует отметить также следующую региональную особенность. В балтийском массиве дискриминативной оказалась лишь верхняя часть распределения по доходам (пятый квинтиль доходов, легковой автомобиль). В российско-казахстанском массиве дискриминативность обоих полюсов доходов оказалась сопоставимой, что является проявлением более низкого уровня жизни в этих регионах.

Утверждения, составившие восьмую гипотезу, подтвердились лишь отчасти. Прежде всего, следует отметить асимметричный характер эффектов социальных страт. Значимо более частая представленность респондентов в кластерах на одном полюсе пространства возможностей вовсе не сопровождалась более редкой представленностью в кластерах на другом полю-

се. Руководители, как и предполагалось, с большей вероятностью вошли в кластеры «успешных». Доля конторских служащих в кластерах «успешных» оказалась существенно выше по сравнению с референтным кластером только в балтийском массиве; в российско-казахстанском массиве этот эффект не был обнаружен. Отмечена значимость лишь одного регрессионного коэффициента для страты работников системы обслуживания, и лишь в балтийском массиве он совпал с предсказанным, хотя знаки других коэффициентов в основном соответствовали предполагавшимся. Предсказания относительно страты профессионалов не подтвердились. Паттерн, предсказанный для этой страты, в российско-казахстанском массиве обнаружен в группе полупрофессионалов. Запутанность общей картины влияния страты на субъективное благосостояние при высокой объяснительной силе этих переменных по коэффициенту псевдо-R-квадрат требует дополнительных исследований в области измерения, кодировки и выбора референтной группы, которые могут быть проведены в рамках предложенной теоретической модели.

Гипотеза о влиянии сектора экономики на самооценку возможностей подтвердилась частично. Для балтийского массива вероятность отнесения к кластеру «неуспешных» в действительности оказалась значимо ниже для работников частного и госбюджетного секторов по сравнению с сектором, финансируемым государством. Для российско-казахстанского массива вероятность попадания в два кластера «успешных» оказалась выше для работающих в частном секторе. Связь между работой в госбюджетном секторе и низкой вероятностью попадания в группу «неуспешных» здесь не наблюдается. Мы объясняем это тем, что в 1993 г., когда проводился опрос, сектор, финансируемый государством (референтный), в России и Казахстане еще не испытывал острых проблем и гарантированная работа в госбюджетном секторе не являлась заметным преимуществом, как в Балтийских странах.

Мы получили полное или частичное подтверждение всех выдвинутых содержательных гипотез, кроме гипотезы о влиянии социальной страты на оценки молодежью своих возможностей, которая нуждается в дополнительной спецификации.

Выводы

Концептуальный синтез теории сравнения и теории условий жизни оказался эвристичным в исследованиях субъективного благополучия населения постсоветских обществ. Оценка своих возможностей молодежью успешно объясняется и предсказывается социально-демографическими (пол, семейный статус, наличие детей, место проживания) и социально-экономическими (образование, занятость, доходы, социальная страта и

сектор экономики) факторами. Темпы рыночных преобразований существенно опосредуют эти влияния, что показали проверка содержательных гипотез о детерминантах субъективного благополучия в переходный период и — особенно убедительно — сравнение регионов с разной скоростью рыночных преобразований.

Литература

1. *Зубова Л. Г., Ковалева Н. В.* Качество жизни в субъективных оценках населения // Экономические и социальные перемены. Мониторинг общественного мнения. 1994, март—апрель. С. 41—43.

2. *Косова Л. Б.* Удовлетворенность жизнью и интенсивность // Социол. исслед. 1994. №10. С.161—164.

3. Theories of Cognitive Consistency: A Source Book / R. Abelson (ed.) [et al.]. Chicago, 1968.

4. *Aldrich J. H., Nelson F. D.* Linear Probability, Logit and Probit Models. Beverly Hills, CA: Sage Publications, 1984.

5. *Andrews F. M., Withey S. B.* Social Indicators of Well-Being: America's Perceptin of Life Quality. New York: Plenum Press, 1976.

6. *Costa P. M., McCrea R. R.* Influence of Extraversion and Neuroticism on Subjective Well-Being: Happy and Unhappy People // Journal of Personality and Social Psychology. 1980. № 38. P. 668—678.

7. *Diener E.* Subjective Well-Being // Psychological Bulletin. 1984. № 95. P. 542—575.

8. The Relationship between Income and Subjective Well-Being: Relative and Absolute / E. Diener [et. al.] // Social Indicators Research. 1993. № 28. P. 195—223.

9. *Easterlin R. A.* Does Economic Growth Improve the Human Lot: Some Empirical Evidence // Nations and Household in Economic Growth. Palo Alto, CA: Stanford University Press, 1974. P. 98—125.

10. *Evans D. R.* Enhansing Quality of Life in the Population at Large // Social Idicators Research. 1994. № 33. P. 47—88.

11. *Festinger L. A.* Theory of Social Comparison Processes // Human Relations. 1954. № 7. P. 117—140.

12. *George L. K.* Economic Status and Subjective Well-Being: A Review of the Literature and the Agenda for Future Research // Aging, Money, and Life Satisfaction. N.Y.: Springer Verlag, 1992.

13. *Heady B.* An Economic Model of Subjective Well-Being: Integrating Economic and Psychological Theories // Social Indicators Research.1993. № 28. P. 97—116.

14. *Juster F. M., Stafford F.P.* Time, Goods and Well-Being. Ann-Arbor: ISR, 1985.

15. *Keller R. T.* Cross-Cultural Influence on Work and Nonwork Contributors to Quality of Life // Group and Organization Studies. 1987. № 12. P. 304—318.

16. *Lance C. E., Mallard A. G., Michalos A. C.* Tests of the Causal Directions of Global-Life Facet Satisfaction Relationship // Social Indicators Research. 1995. № 34. P. 69—92.

17. *Lazarus R. S.* Cognition and Motivation in Emotion // American Psychologist. 1991. № 46. P. 352—367.

18. *Michalos A. C.* Global Report on Student Well-Being. Life Satisfaction and Happiness. Vol. 1. N.Y.: Springer Verlag, 1991.

19. *Schulz W.* Multiple-Descrepancies Theory Versus Resource Theory // Social Indicators research. 1995. № 34. P. 153–169.

20. *Tuma N. B.* Invoking RATE. (Manuscript). 1992.

21. *Veenhoven R.* Is Happiness Relative? Social Indicators Research. 1991. № 24. P. 1–34.

22. *Veenhoven R.* The Cross-National Pattern of Happiness: Test of Predictions Implied in Three Theories of Happiness // Social Indicators Research. 1995. № 34. P. 33–68.

23. *Zanna M. P., Olson J. M., Fazio R. H.* Self-Perception and Attitude-Behavior Consistency // Personality and Psychology Bulletin. 1981. № 7. P. 252–256.

ОГЛАВЛЕНИЕ

Учебное издание

Классическое университетское издание

Терещенко Ольга Викентовна
Курилович Наталия Вячеславовна
Князева Екатерина Ивановна

МНОГОМЕРНЫЙ СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ В СОЦИАЛЬНЫХ НАУКАХ

Учебное пособие

Редактор *Г. В. Лозовская*
Художник обложки *Т. Ю. Таран*
Технический редактор *Т. К. Раманович*
Компьютерная верстка *О. В. Гасюк*
Корректор *А. А. Заяш*

Подписано в печать 29.12.2012. Формат 60×90/16. Бумага офсетная.
Печать офсетная. Усл. печ. л. 15,0. Уч.-изд. л. 17,13.
Тираж 150 экз. Заказ 169.

Белорусский государственный университет.
ЛИ № 02330/0494425 от 08.04.2009.
Пр. Независимости, 4, 220030, Минск.

Республиканское унитарное предприятие
«Издательский центр Белорусского государственного университета».
ЛП № 02330/0494178 от 03.04.2009.
Ул. Красноармейская, 6, 220030, Минск.